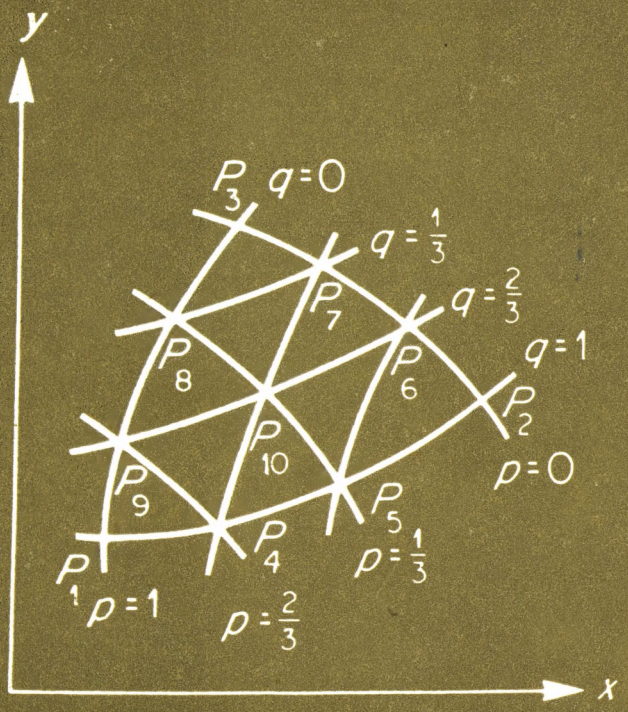
Э. Митчелл, Р. Уэйт



МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ


ИЗДАТЕЛЬСТВО МИР

THE FINITE ELEMENT METHOD IN PARTIAL DIFFERENTIAL EQUATIONS

A. R. MITCHELL
Department of Mathematics, University of Dundee

and

R. WAIT
Department of Computational aud Statistical Science,
University of Liverpool

A Wiley — Interscience Publication

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto

Э. Митчелл
Р. Уэйт

МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Перевод с английского
В. Е. КОНДРАШОВА и В. Ф. КУРЯКИНА

под редакцией
Н. Н. ЯНЕНКО

Издательство «Мир» Москва 1981

УДК 519.3

Предлагаемая книга посвящена методу конечных элементов и отличается от других книг по этой тематике простотой и компактностью изложения, широтой охвата материала и методичностью изложения. В книге даются анализ различных вариантов метода и многочисленные примеры его применения к конкретным задачам. Приведено свыше ста упражнений различной степени трудности.

Книга полезна для специалистов, применяющих метод конечных элементов на практике, и студентов, специализирующихся в области прикладной математики.

Редакция литературы по математическим наукам

1702070000

М $\frac{20204\text{-}027}{041(01)\text{-}81}$ 27-81, ч. 1

ПРЕДИСЛОВИЕ РЕДАКТОРА ПЕРЕВОДА

Метод конечных элементов, который начал интенсивно разрабатываться с середины 60-х годов, стал теперь достаточно эффективным способом численного решения целого ряда задач для уравнений в частных производных, в особенности для эллиптических нестационарных уравнений. Он очень удобен для программирования и позволяет учитывать дополнительную информацию о решаемой задаче в тех случаях, когда удается получить теоретическое обоснование его применимости.

Многое еще предстоит сделать для совершенствования этого метода и расширения сферы его применения — прежде всего к нестационарным нелинейным задачам, для которых конечно-разностный метод остается пока основным способом получения численных решений. Но достигнутые уже сейчас уровень теоретической обоснованности и широта практических приложений метода конечных элементов делают весьма желательным обучение будущих специалистов по прикладной математике основам этого метода.

В нашей стране уже вышло немало книг, посвященных методу конечных элементов, в том числе и переводы трудов ведущих зарубежных ученых, но все это либо монографии для специалистов, либо учебные пособия для инженеров. Авторы настоящей книги предприняли одну из первых и, как нам кажется, весьма успешную попытку создать учебное руководство для студентов, обучающихся прикладной математике, и практических работников вычислительных центров, не знакомых еще с этим методом.

Прочитав книгу и, в особенности, решив хотя бы часть приведенных в ней задач, читатель приобретет определенные навыки проведения подготовительной работы, необходимой при решении конкретных задач методом конечных элементов,

а также получит достаточно ясное представление о теоретических основах метода. Немало интересного найдут в книге и специалисты — большой набор базисных функций, сравнительный анализ различных вариантов метода конечных элементов; большое внимание авторы уделяют применению метода для решения нестационарных задач.

Книга написана просто и ясно, на хорошем математическом уровне. В ней достаточно полно отражено то большое влияние, которое оказали на обоснование и развитие метода конечных элементов работы советских математиков. Дополнительная библиография поможет читателю получить об этом более детальное представление, а также лучше понять роль и место метода конечных элементов в прикладной математике.

При переводе были исправлены замеченные опечатки и мелкие погрешности, в библиографии некоторые зарубежные работы снабжены ссылками на их русские переводы, а переводы советских работ заменены оригиналами.

Н. Н. Яненко

20 апреля 1979 г.

Энн и Фионе

ПРЕДИСЛОВИЕ

Можно считать общепризнанным, что метод конечных элементов является эффективным способом численного решения дифференциальных уравнений с частными производными. Это в особенности верно для эллиптических уравнений, где сразу проявились его преимущества по сравнению с конечно-разностным методом. Метод конечных элементов служит хорошим примером весьма трудной темы, развитие которой стало возможным только благодаря тесному сотрудничеству между инженерами, математиками и специалистами по численному анализу. Принимая во внимание широту интересов его приверженцев, нетрудно понять, почему по методу конечных элементов не написано книги, которая отражала бы должным образом все возрастающий поток публикаций, ему посвященных. Целью нашей книги было заполнить пробел между хорошо известными работами Зенкевича (1975) и Стренга и Фикса (1977), в которых соответственно нашли отражение запросы инженеров и математиков. В старинном споре о сравнительных преимуществах методов конечных разностей и конечных элементов мы не становимся ни на одну сторону — нас вполне удовлетворяет, что есть два таких мощных метода численного решения дифференциальных уравнений с частными производными.

Большая часть книги доступна студентам соответствующих математических и инженерных специальностей. Для ее понимания не требуется специальных математических знаний, выходящих за рамки обычных курсов линейной алгебры и анализа. Исключением является гл. 5, которую при первом чтении можно опустить. Гильбертово пространство и понятия из функционального анализа используются на протяжении всей книги главным образом для унификации изложения материала. Но мы предполагаем, что у читателя есть определенные навыки практической работы с дифференциальными уравнениями в частных производных — только в этом случае наша книга будет для него действительно полезной. Так как отправной точкой для нас чаще являются не уравнения с частными производными, а тот или иной вариационный принцип, то в книгу включена глава о вариационных принципах с детальными ссылками на более подробные руководства.

Мы надеемся, что книга окажется полезной и для специалистов, применяющих метод конечных элементов на практике. С учетом их интересов метод конечных элементов излагается в самых различных вариантах (а именно: в форме Ритца, Галеркина, наименьших квадратов, коллокации), а в гл. 4 приводится большой набор возможных базисных функций, которые могут быть использованы в каждом из этих вариантов. Чтобы сбалансировать включение такого большого по объему практического материала, мы полностью опустили задачи на собственные значения. Некоторым оправданием этого шага служит то, что задачи на собственные значения в близкой нам трактовке подробно изложены в гл. 6 книги Стренга и Фикса (1977).

Библиография была ограничена только теми работами, на которые в книге делаются ссылки по существу. Более полная библиография по методу конечных элементов имеется в книге Уайтмена (1975). Учитывая возможный интерес читателя, мы включили в наш список литературы последние монографии и труды конференций, посвященные главным образом методу конечных элементов (см. Зенкевич (1975), Азиз (1972), Оден (1976), Стренг и Фикс (1977), Грам (1973), Ланкастер (1973, 1975), Миллер (1973, 1975), Уайтмен (1973, 1976), Уотсон (1974, 1976), де Бур (1974), Оден, Зенкевич, Галлагер и Тейлор (1974), Прентер (1975), Хаббард (1971)).

Большая часть материала этой книги излагалась в виде лекций для аспирантов и студентов математических специальностей в университетах Данди и Ливерпуля. Кроме того, первый автор по приглашению Института атомной энергии читал лекции по материалу гл. 2 и 4 в Институте высших исследований НАТО в Кьелере (Норвегия) в 1973 г., а второй — лекции по материалу гл. 3, 5 и 6 в 1973 г. во время своего пребывания в Техническом университете Дании.

Мы весьма признательны нашим коллегам и бывшим студентам за те полезные обсуждения, которые проводились в процессе подготовки этой книги. Мы особенно благодарны Бобу Барнхиллу, Лотару Коллатцу, Дэвиду Гриффитсу, Дирку Лаури, Джеку Ламберту, Петеру Ланкастеру, Робину Маклеоду, Гилу Стренгу, Юджину Уачспрессу, Джиму Уатту и Олеку Зенкевичу. И, наконец, мы благодарны Роз Даджон и Дорин Мэнли, которые с большой тщательностью перепечатали рукопись.

ГЛАВА 1

ВВЕДЕНИЕ

1.1. Аппроксимация кусочно-полиномиальными функциями

Рассмотрим вначале задачу об аппроксимации вещественной функции $f(x)$ на конечном интервале оси x. Один из простых способов решения этой задачи состоит в разбиении интервала на некоторое число неперекрывающихся подынтервалов и линейной интерполяции по значениям функции $f(x)$ в граничных точках подынтервалов (см. рис. 1(а)). Если имеется n подынтервалов $[x_i, x_{i+1}]$ $(i = 0, 1, 2, \ldots, n-1)$, то кусочно-линейная аппроксимирующая функция зависит только от значений функции $f_i (= f(x_i))$ в узловых точках x_i $(i = 0, 1, 2, \ldots, n)$. В тех задачах, где $f(x)$ задается неявно уравнением (дифференциальным, интегральным, функциональным и т. д.), значения f_i являются неизвестными параметрами задачи. В задаче интерполяции значения f_i известны заранее.

На подынтервале $[x_i, x_{i+1}]$ соответствующая часть аппроксимирующей функции описывается формулой

$$p_1^{(i)}(x) = \alpha_i(x) f_i + \beta_{i+1}(x) f_{i+1} \quad (x_i \leqslant x \leqslant x_{i+1}), \tag{1.1}$$

где

$$\alpha_i(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i} \quad \text{и} \quad \beta_{i+1}(x) = \frac{x - x_i}{x_{i+1} - x_i} \quad (i = 0, 1, 2, \ldots, n-1).$$

Следовательно, кусочная аппроксимирующая функция на интервале $x_0 \leqslant x \leqslant x_n$ задается формулой

$$p_1(x) = \sum_{i=0}^{n} \varphi_i(x) f_i, \tag{1.2}$$

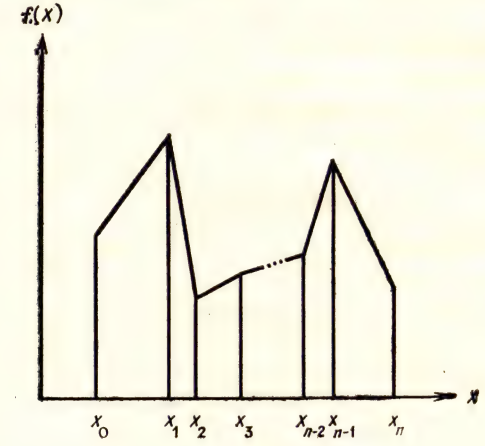
где

$$\varphi_0(x) = \begin{cases} \dfrac{x_1 - x}{x_1 - x_0} & (x_0 \leqslant x \leqslant x_1) \\ 0 & (x_1 \leqslant x \leqslant x_n), \end{cases}$$

$$\varphi_i(x) = \begin{cases} 0 & (x_0 \leqslant x \leqslant x_{i-1}) \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & (x_{i-1} \leqslant x \leqslant x_i) \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & (x_i \leqslant x \leqslant x_{i+1}) \\ 0 & (x_{i+1} \leqslant x \leqslant x_n), \end{cases} \tag{1.3}$$

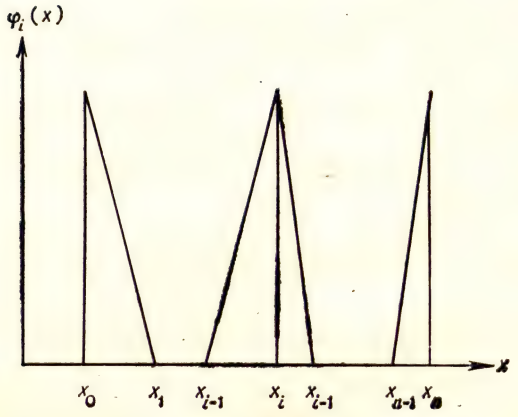
$$\varphi_n(x) = \begin{cases} 0 & (x_0 \leqslant x \leqslant x_{n-1}) \\ \dfrac{x - x_{n-1}}{x_n - x_{n-1}} & (x_{n-1} \leqslant x \leqslant x_n) \end{cases}$$

являются пирамидальными функциями, изображенными на рис. 1(b). Пирамидальные функции, заданные формулами (1.3), представляют простейший тип базисных функций. Отметим, в частности, что базисные функции $\varphi_i(x)$ $(i = 1, 2, \ldots$



a

Рис. 1 (a).



b

Рис. 1 (b).

..., $n-1$) равны нулю вне интервала $[x_{i-1}, x_{i+1}]$ или, как говорят, имеют локальный носитель. В этой книге будут строиться базисные функции различной степени сложности, однако всегда с локальными носителями. Основным свойством большинства базисных функций является то, что они равны единице в некоторой узловой точке и равны нулю в большинстве других узловых точек.

Вообще говоря, первые производные кусочно-полиномиальной аппроксимирующей функции $p_1(x)$, заданной формулой (1.2), отличаются от первой производной $f(x)$ даже в узлах. Теперь попытаемся построить кусочную аппроксимирующую функцию, которая совпадала бы вместе с первой производной с $f(x)$ в узловых точках x_i ($i=0, 1, 2, \ldots, n$). Другими словами, мы должны построить кусочно-кубический полином $p_3(x)$, такой, что

$$D^k f(x_i) = D^k p_3(x_i) \quad (k=0, 1;\ i=0, 1, 2, \ldots, n),$$

где $D = d/dx$. На подынтервале $[x_i, x_{i+1}]$ соответствующая часть кубического аппроксимирующего полинома задается формулой

$$p_3^{(i)}(x) = \alpha_i(x) f_i + \beta_{i+1}(x) f_{i+1} + \gamma_i(x) f_i' + \delta_{i+1}(x) f_{i+1}', \quad (1.4)$$

где

$$\alpha_i(x) = \frac{(x_{i+1}-x)^2[(x_{i+1}-x_i)+2(x-x_i)]}{(x_{i+1}-x_i)^3},$$

$$\beta_{i+1}(x) = \frac{(x-x_i)^2[(x_{i+1}-x_i)+2(x_{i+1}-x)]}{(x_{i+1}-x_i)^3},$$

$$\gamma_i(x) = \frac{(x-x_i)(x_{i+1}-x)^2}{(x_{i+1}-x_i)^2}, \quad (1.5)$$

$$\delta_{i+1}(x) = \frac{(x-x_i)^2(x-x_{i+1})}{(x_{i+1}-x_i)^2}$$

($i=0\ 1, 2, \ldots, n-1$) и штрих означает дифференцирование по x. Кусочная аппроксимирующая функция на интервале $x_0 \leqslant x \leqslant x_n$ задается формулой

$$p_3(x) = \sum_{i=0}^{n} [\varphi_i^{(0)}(x) f_i + \varphi_i^{(1)}(x) f_i'], \quad (1.6)$$

где кубические полиномы $\varphi_i^{(0)}(x)$, $\varphi_i^{(1)}(x)$ ($i=0, 1, 2, \ldots, n$) легко получаются из (1.5). Базисные функции $\varphi_i^{(0)}(x)$ и $\varphi_i^{(1)}(x)$ ($i=1, 2, \ldots, n-1$) изображены на рис. 2.

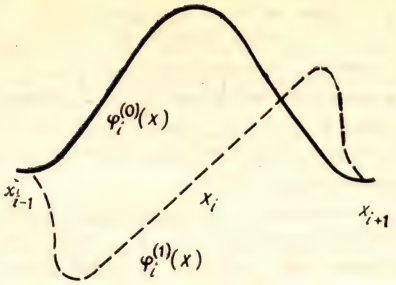


Рис. 2.

Базисные функции в выражениях (1.2) и (1.6) появляются при рассмотрении частных случаев *кусочной эрмитовой интерполяции* (или аппроксимации) для заданного разбиения интервала. Пусть теперь в общем случае Π: $a=x_0<x_1<\ldots$ $\ldots<x_n=b$ есть произвольное разбиение интервала $R=[a, b]$ на оси x. Для целого положительного m и разбиения интервала Π обозначим через $H=H^{(m)}(\Pi, R)$ множество всех действительных кусочно-полиномиальных функций $w(x)$, определенных на R, так, что $w(x)\in C^{m-1}(R)$ и $w(x)$ есть полином степени $2m-1$ на каждом подынтервале $[x_i, x_{i+1}]$ интервала R. Для любой заданной действительной функции $f(x)\in C^{m-1}(R)$ кусочная эрмитова интерполяция однозначно определяется как такой элемент $p_{2m-1}(x)\in H$, для которого

$$D^k f(x_i)=D^k p_{2m-1}(x_i)\begin{cases}(0\leqslant k\leqslant m-1)\\(0\leqslant i\leqslant n).\end{cases} \tag{1.7}$$

Частные случаи при $m=1$, 2 уже были рассмотрены выше и привели к базисным функциям, входящим в выражения (1.2) и (1.6) соответственно. Оценки ошибок для кусочных эрмитовых приближений приводятся в работе Биркгофа, Шульца и Варги (1968).

В задачах, где требуется определить только $f(x)$, часто бывает нежелательно вводить в качестве дополнительных параметров производные $f(x)$ и тем самым заметно увеличивать порядок системы уравнений, которая должна решаться. Поэтому весьма желательным свойством кусочных функций является непрерывность производных в точках сшивки полиномов без введения значений производных в качестве дополнительных неизвестных параметров. Простейшим примером такого подхода представляется подбор на каждом подынтервале $[x_i, x_{i+1}]$ $(i=0, 1, 2, \ldots, n-1)$ такой параболы, чтобы первые производные были непрерывны в каждой внутренней узловой точке x_i $(i=1, 2, \ldots, n-1)$. Удобную форму такой

кусочной аппроксимации представляет квадратичный *сплайн*

$$S_2^{(i)}(x) = f_i + \frac{f_{i+1} - f_i}{x_{i+1} - x_i}(x - x_i) + c_i(x - x_i)(x - x_{i+1}) \quad (1.8)$$
$$(i = 0, 1, 2, \ldots, n-1),$$

где из непрерывности первых производных следует, что

$$c_i + c_{i-1} = \frac{1}{h^2}(f_{i+1} - 2f_i + f_{i-1}) \quad (i = 1, 2, \ldots, n-1), \quad (1.9)$$

а разбиение интервала предполагается равномерным с шагом h. Система (1.9) представляет собой $n-1$ линейных соотношений относительно неизвестных коэффициентов c_i ($i = 0, 1, 2, \ldots, n-1$), и поэтому в случае квадратичных сплайнов остается свободным один коэффициент. Поскольку $S_2^{(i)''}(x) = 2c_i$ ($i = 0, 1, 2, \ldots, n-1$), знание второй производной в любой точке полностью решает задачу.

Наибольшее признание получил кубический сплайн [1]), для которого при заданных значениях f_i ($i = 0, 1, 2, \ldots, n$) на каждом подынтервале подбираются кубические полиномы, такие, чтобы первая и вторая производные были непрерывны во всех внутренних узловых точках. Если $S_3^{(i)}(x)$ ($i = 0, 1, 2 \ldots \ldots, n-1$) есть искомый кубический сплайн, то функция $S_3^{(i)''}(x)$ должна быть линейной на $[x_i, x_{i+1}]$, и поэтому

$$S_3^{(i)''}(x) = c_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + c_{i+1} \frac{x - x_i}{x_{i+1} - x_i} \quad (i = 0, 1, 2, \ldots, n-1),$$

где c_i, c_{i+1} являются значениями вторых производных в точках x_i, x_{i+1} соответственно. Это обеспечивает непрерывность вторых производных во внутренних узловых точках. Используя дополнительные условия

$$\left.\begin{aligned} S_3^{(i)}(x_i) &= f_i \\ S_3^{(i)}(x_{i+1}) &= f_{i+1} \end{aligned}\right\} \quad (i = 0, 1, 2, \ldots, n-1)$$

и

$$S_3^{(i-1)'}(x_i) = S_3^{(i)'}(x_i) \quad (i = 1, 2, \ldots, n-1),$$

---

[1]) При $h \to 0$ кубический сплайн равномерно сходится к аппроксимируемой функции при условии достаточной ее гладкости. Квадратичные сплайны таким свойством не обладают. Если, например, на левом конце интервала задать вторую производную, то по мере приближения к правому концу эти сплайны начинают осциллировать, тем сильнее, чем меньше h. — *Прим. перев.*

получим кубический сплайн в виде

$$S_3^{(i)}(x) = \frac{c_i}{6h}(x_{i+1} - x)^3 + \frac{c_{i+1}}{6h}(x - x_i)^3 + \\ + \left(\frac{f_i}{h} - \frac{hc_i}{6}\right)(x_{i+1} - x) + \left(\frac{f_{i+1}}{h} - \frac{hc_{i+1}}{6}\right)(x - x_i) \quad (1.10)$$
$$(i = 0, 1, 2, \ldots, n-1),$$

где сетка предполагается равномерной, а $n+1$ коэффициентов c_i $[i = 0, 1, 2, \ldots, n)$ удовлетворяют системе $n-1$ линейных соотношений

$$c_{i+1} + 4c_i + c_{i-1} = \frac{6}{h^2}(f_{i+1} - 2f_i + f_{i-1}) \quad (i = 1, 2, \ldots, n-1). \quad (1.11)$$

Два свободных параметра в случае кубического сплайна часто исключают, полагая $c_0 = c_n = 0$, и поэтому другие параметры однозначно определяются из (1.11).

Более естественным представлением кубического сплайна при равномерном разбиении интервала $I = [0, b]$ является выражение

$$S_I\left(\frac{x}{h}\right) = \alpha_0 + \alpha_1\left(\frac{x}{h}\right) + \alpha_2\left(\frac{x}{h}\right)^2 + \alpha_3\left(\frac{x}{h}\right)^3 + \\ + \sum_{s=1}^{n-1} \beta_s \left(\frac{x}{h} - s\right)_+^3, \quad (1.12)$$

где

$$\left(\frac{x}{h} - s\right)_+ = \begin{cases} 0 & \left(\frac{x}{h} \leqslant s\right) \\ \frac{x}{h} - s & \left(\frac{x}{h} > s\right). \end{cases}$$

Нетрудно показать, что функция $S_I(x/h)$ и все ее производные, кроме третьей, непрерывны во всех внутренних узловых точках x_i $(i = 1, 2, \ldots, n-1)$ при всех значениях $n+3$ коэффициентов α_0, α_1, α_2, α_3, β_s $(s = 1, 2, \ldots, n-1)$. Условие

$$S_I\left(\frac{x_i}{h}\right) = f_i \quad (i = 0, 1, 2, \ldots, n), \quad (1.13)$$

дает $n+1$ линейных соотношений для $n+3$ коэффициентов, и поэтому остаются два свободных параметра. Система линейных уравнений сводится к виду, приведенному в упражнении 4. Если кубический сплайн (1.12) вместе с двумя свобод-

ными параметрами теперь представить в виде

$$S_I\left(\frac{x}{h}\right)=\sum_{i=0}^{n} f_i C_i\left(\frac{x}{h}\right), \tag{1.14}$$

где $C_i(x_i/h)=1$, $C_i(x_j/h)=0 (j\neq i)$, $(i, j=0, 1, 2, \ldots, n)$, то получается, что основные сплайны $C_i(x/h)$ не имеют локального носителя и не являются удобными для практики базисными функциями.

Кубические сплайны с локальным носителем длины $4h$ были предложены Шёнбергом (1969) как удобные базисные функции. Для узловых точек $x=ih$ $(i=2, 3, \ldots, n-2)$ они имеют вид

$$B_i\left(\frac{x}{h}\right)=\frac{1}{4}\left[\left\{\frac{x}{h}-(i-2)\right\}_+^3-4\left\{\frac{x}{h}-(i-1)\right\}_+^3+\right.$$
$$\left.+6\left\{\frac{x}{h}-i\right\}_+^3-4\left\{\frac{x}{h}-(i+1)\right\}_+^3+\left\{\frac{x}{h}-(i+2)\right\}_+^3\right]. \tag{1.15}$$

Эти функции и две их первые производные равны нулю при $-\infty < x/h \leqslant i-2$ и $i+2 \leqslant x/h < +\infty$. Кроме того,

$$B_i(i-1)=B_i(i+1)=\frac{1}{4}, \quad B_i(i)=1 \quad (i=2, 3, \ldots, n-2).$$

Остальные функции $B_0(x/h)$, $B_1(x/h)$, $B_{n-1}(x/h)$, $B_n(x/h)$ требуют специального рассмотрения. Полагая

$$S_I\left(\frac{x}{h}\right)=\sum_{i=0}^{n} \gamma_i B_i\left(\frac{x}{h}\right) \tag{1.16}$$

и сопоставляя правые части (1.14) и (1.16), мы получим трехдиагональную систему линейных уравнений для определения коэффициентов γ_i $(i=0, 1, 2, \ldots, n)$, входящих в (1.16). Большинство уравнений этой системы имеет вид

$$\gamma_i+\frac{1}{4}(\gamma_{i+1}+\gamma_{i-1})=f_i \quad (i=2, 3, \ldots, n-2).$$

Двумерная аппроксимация

Теперь рассмотрим задачу об аппроксимации действительной функции кусочными непрерывными функциями на ограниченной области R с границей ∂R. Область разбивается на некоторое число элементов. В этом разделе рассматриваются области, имеющие вид прямоугольника или многоугольника.

1) *Прямоугольная область.* Стороны такой области параллельны осям x и y, и она разбивается на такие же прямоуголь-

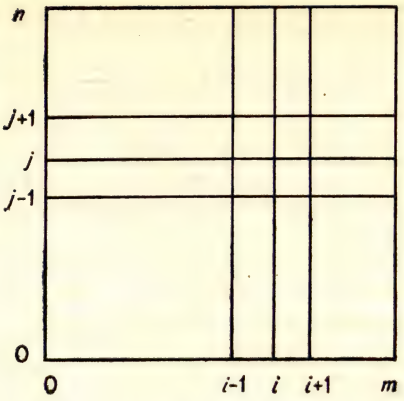


Рис. 3.

ные элементы прямыми линиями, параллельными осям. Пусть $[x_0, x_m] \times [y_0, y_n]$ есть прямоугольная область и $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ есть типичный прямоугольный элемент, где $x_{i+1} - x_i = h_1$ и $y_{j+1} - y_j = h_2$ $(0 \leqslant i \leqslant m-1,\ 0 \leqslant j \leqslant n-1)$ (см. рис. 3). Билинейная форма, приближающая функцию $f(x, y)$ на прямоугольном элементе, имеет вид

$$p_1^{(i,j)}(x, y) = \alpha_{i,j}(x, y) f_{i,j} + \beta_{i+1,j}(x, y) f_{i+1,j} + \gamma_{i,j+1}(x, y) f_{i,j+1} + \delta_{i+1,j+1}(x, y) f_{i+1,j+1}, \quad (1.17)$$

где

$$\alpha_{i,j}(x, y) = \frac{1}{h_1 h_2}(x_{i+1} - x)(y_{j+1} - y),$$

$$\beta_{i+1,j}(x, y) = \frac{1}{h_1 h_2}(x - x_i)(y_{j+1} - y),$$

$$\gamma_{i,j+1}(x, y) = \frac{1}{h_1 h_2}(x_{i+1} - x)(y - y_j)$$

и

$$\delta_{i+1,j+1}(x, y) = \frac{1}{h_1 h_2}(x - x_i)(y - y_j)$$

$(0 \leqslant i \leqslant m-1;\ 0 \leqslant j \leqslant n-1)$. Кусочная аппроксимирующая функция в области $[x_0, x_m] \times [y_0, y_n]$ задается выражением

$$p_1(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{n} \varphi_{i,j}(x, y) f_{i,j}. \quad (1.18)$$

Базисные функции $\varphi_{i,j}(x, y)$ $(1 \leqslant i \leqslant m-1;\ 1 \leqslant j \leqslant n-1)$ обращаются в нуль всюду, за исключением прямоугольной области $[x_{i-1}, x_{i+1}] \times [y_{j-1}, y_{j+1}]$, и поэтому имеют локальный носитель (см. упражнение 6 и рис. 3).

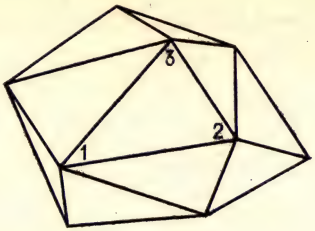


Рис. 4.

Только что рассмотренный случай является простейшим примером кусочной двумерной эрмитовой интерполяции (или аппроксимации) на прямоугольной области, разбитой на прямоугольные элементы. В общем случае для любого целого положительного числа l и любого разбиения прямоугольника R на прямоугольные элементы обозначим через $H = H^{(l)}(R)$ совокупность всех действительных кусочных полиномов $g(x, y)$, определенных на R, так, что $g(x, y) \in C^{l-1, l-1}(R)$ и $g(x, y)$ есть полином степени $2l - 1$ по каждому переменному x и y на каждом прямоугольном элементе $[x_i, x_{i+1}] \times (y_j, y_{j+1}]$ $(0 \leqslant i \leqslant m - 1;\ 0 \leqslant j \leqslant n - 1)$ области R. Для любой заданной действительной функции $f(x, y) \in C^{l-1, l-1}(R)$ существует единственный кусочный эрмитов интерполянт $p_{2l-1}(x, y) \in H$, определяемый условиями

$$D^{(p, q)} f(x_i, y_j) = D^{(p, q)} p_{2l-1}(x_i, y_j)$$

при всех $0 \leqslant p, q \leqslant l - 1$, $0 \leqslant i \leqslant m - 1$, $0 \leqslant j \leqslant n - 1$. Частный случай $l = 1$ был уже рассмотрен выше и дает билинейные базисные функции, вид которых указан в упражнении 6. Случай $l = 2$ разбирается в упражнении 7. Читателя, интересующегося получением оценок ошибок для двумерной эрмитовой интерполяции, мы снова отсылаем к работе Биркгофа, Шульца и Варги (1968).

2) *Многоугольная область.* Под этим понимается либо область в форме многоугольника, либо аппроксимация области другой формы. Многоугольник произвольным образом разбивается на треугольные элементы. Для типичного треугольного элемента с вершинами (x_i, y_i) $(i = 1, 2, 3)$ (см. рис. 4) линейная форма, приближающая функцию $f(x, y)$ на треугольном элементе, имеет вид

$$p_1(x, y) = \sum_{i=1}^{3} \alpha_i(x, y) f_i, \tag{1.19}$$

где $f_i = f(x_i, y_i)$ $(i = 1, 2, 3)$. Коэффициенты $\alpha_i(x, y)$ $(i = 1, 2, 3)$ задаются формулами

$$\alpha_1(x, y) = \frac{1}{C_{123}}(\tau_{23} + \eta_{23}x - \xi_{23}y),$$
$$\alpha_2(x, y) = \frac{1}{C_{123}}(\tau_{31} + \eta_{31}x - \xi_{31}y), \qquad (1.20)$$
$$\alpha_3(x, y) = \frac{1}{C_{123}}(\tau_{12} + \eta_{12}x - \xi_{12}y),$$

где $|C_{123}|$ есть удвоенная площадь треугольника, а

$$\tau_{ij} = x_i y_j - x_j y_i,$$
$$\xi_{ij} = x_i - x_j,$$
$$\eta_{ij} = y_i - y_j \quad (i, j = 1, 2, 3).$$

Разумеется, функции (1.20) являются только частями полных базисных функций, связанных с вершинами треугольной сетки. Полная базисная функция относительно некоторой вершины получается путем суммирования частей, связанных с теми треугольниками, которые примыкают к этой вершине. Например, вершина 1 на рис. 4 имеет пять примыкающих треугольников, и поэтому базисная функция, соответствующая этой вершине, будет состоять из пяти частей. Полная базисная функция оказывается пирамидальной.

Упражнение 1. Покажите, что кубический полином $p_3(x)$, который принимает значения

$$p_3(0) = f_0, \quad p_3(1) = f_1, \quad p_3'(0) = f_0', \quad p_3'(1) = f_1',$$

определяется формулой

$$p_3(x) = (1 - x)^2(1 + 2x) f_0 + x(1 - x)^2 f_0' + x^2(3 - 2x) f_1 + x^2(x - 1) f_1'.$$

Упражнение 2. Используйте результат упражнения 1, чтобы получить коэффициенты в представлении (1.4) и тем самым получить базисные функции, входящие в (1.6).

Упражнение 3. Применяя намеченный в основном тексте метод, получите выражение для кубического сплайна в форме (1.10), где коэффициенты определяются соотношениями (1.11).

Упражнение 4. Применяя условие (1.13) к сплайну вида (1.12), покажите, что система уравнений для определения

входящих в (1.12) коэффициентов имеет вид

$$\begin{aligned}
\beta_{i-1}+4\beta_i+\beta_{i+1}&=\delta^4 f_i \quad (i=2,\ 3,\ \ldots,\ n-2),\\
6\alpha_3+5\beta_1+\beta_2&=\delta^3 f_{3/2},\\
2\alpha_2+6\alpha_3+\beta_2&=\delta^2 f_1,\\
\alpha_1+\alpha_2+\alpha_3&=\delta f_{1/2},\\
\alpha_0&=f_0,
\end{aligned}$$

где δ есть обычный оператор взятия центральной разности.

Упражнение 5. Решите систему уравнений упражнения 4 при $\alpha_1=\alpha_2=0$ и $n=4$. Покажите с помощью (1.14), что в этом случае

$$C_2\left(\frac{x}{h}\right)=\left(\frac{x}{h}-1\right)_+^3-8\left(\frac{x}{h}-2\right)_+^3+37\left(\frac{x}{h}-3\right)_+^3.$$

Нарисуйте приближенный график основного сплайна $C_2(x/h)$ при $-\infty<x/h<+\infty$ и покажите, что его носитель не является локальным.

Упражнение 6. Покажите, что для единичного квадрата при $l=1$ базисные функции, соответствующие внутренним узлам, имеют вид

$$\varphi_{i,j}(x,\ y)=\begin{cases}
\left[\frac{x}{h}-(i-1)\right]\left[\frac{y}{h}-(j-1)\right] & \left(i-1\leqslant\frac{x}{h}\leqslant i;\ j-1\leqslant\frac{y}{h}\leqslant j\right)\\
\left[\frac{x}{h}-(i-1)\right]\left[(j+1)-\frac{y}{h}\right] & \left(i-1\leqslant\frac{x}{h}\leqslant i;\ j\leqslant\frac{y}{h}\leqslant j+1\right)\\
\left[(i+1)-\frac{x}{h}\right]\left[\frac{y}{h}-(j-1)\right] & \left(i\leqslant\frac{x}{h}\leqslant i+1;\ j-1\leqslant\frac{y}{h}\leqslant j\right)\\
\left[(i+1)-\frac{x}{h}\right]\left[(j+1)-\frac{y}{h}\right] & \left(i\leqslant\frac{x}{h}\leqslant i+1;\ j\leqslant\frac{y}{h}\leqslant j+1\right)\\
0 & \text{при всех других значениях аргументов,}
\end{cases}$$

где $1\leqslant i,\ j\leqslant m-1$ и $mh=1$.

Упражнение 7. На единичном квадрате рассмотрите полином

$$g(x, y) = \sum_{r=0}^{3} \sum_{s=0}^{3} \alpha_{rs} x^r y^s.$$

Выразите все коэффициенты α_{rs} $(0 \leqslant r, s \leqslant 3)$ через значения функций g, $\partial g/\partial x$, $\partial g/\partial y$ и $\partial^2 g/\partial x \partial y$ в четырех угловых точках квадрата. Покажите, что результаты этих вычислений могут быть использованы для нахождения базисных функций в случае $l = 2$ общей теории двумерной эрмитовой интерполяции на прямоугольной области, разбитой на прямоугольные элементы.

1.2. Функциональные пространства

Этот параграф содержит введение в математические структуры, необходимые для понимания теоретических аспектов метода конечных элементов. Будет изложен только самый необходимый материал, а интересующемуся читателю мы рекомендуем обратиться к книгам Симмонса (1963) или Иосиды (1967).

Линейным или *векторным пространством* называется непустое множество X, в котором любые два элемента x и y могут быть объединены операцией, называемой *сложением*, так что в результате получается некоторый элемент из X, обозначаемый как $x + y$, причем операция сложения удовлетворяет следующим условиям:

(I) $x + y = y + x$,
(II) $x + (y + z) = (x + y) + z$,
(III) существует нулевой элемент φ, такой, что $\varphi + x = x + \varphi = x$ для всех x,
(IV) для каждого x существует отрицательный элемент $-x$, такой, что $x + (-x) = \varphi$.

Еще одна необходимая для линейного пространства операция состоит в том, что любой элемент $x \in X$ может быть объединен с любым действительным числом или *скаляром* α. Операция эта называется умножением на скаляр и ее результат обозначается через αx. Умножение на скаляр должно удовлетворять следующим условиям:

(V) $\alpha(x + y) = \alpha x + \alpha y$,
(VI) $(\alpha + \beta) x = \alpha x + \beta x$,
(VII) $(\alpha\beta) x = \alpha(\beta x)$,
(VIII) $1 \cdot x = x$.

Одним из примеров линейного пространства является множество всех N-мерных действительных векторов, для которых $\mathbf{a}+\mathbf{b}=\mathbf{c}$ определяется как $a_i+b_i=c_i$ $(i=1, 2, \ldots, N)$ и $\alpha\mathbf{a}=\mathbf{d}$ как $d_j=\alpha a_j$ $(j=1, 2, \ldots, N)$.

Нормированное линейное пространство (н. л. п.) есть линейное пространство, для которого определена норма каждого элемента x, обозначаемая через $\|x\|$ и удовлетворяющая следующим условиям:

(I) $\|x\|\geqslant 0$,
(II) $\|x\|=0\Leftrightarrow x=0$,
(III) $\|x+y\|\leqslant\|x\|+\|y\|$,
(IV) $\|\alpha x\|=|\alpha|\,\|x\|$.

Таким образом, появляется понятие длины элемента в линейном пространстве. *Полунорма* удовлетворяет условиям (I), (III) и (IV), но не удовлетворяет условию (II).

Пространство с внутренним или скалярным произведением есть линейное пространство, на котором для каждой пары вектором определена действительная функция (x, y), удовлетворяющая условиям:

(I) $(\alpha x+\beta y, z)=\alpha(x, z)+\beta(y, z)$,
(II) $(x, y)=(y, x)$,
(III) $(x, x)\geqslant 0$; $(x, x)=0\Leftrightarrow x=0$.

Упражнение 8. Покажите, что линейное пространство со скалярным произведением является линейным нормированным пространством относительно нормы

$$\|x\|=(x, x)^{1/2}.$$

Затем убедитесь в справедливости правила параллелограмма

$$\|x+y\|^2+\|x-y\|^2=2\|x\|^2+2\|y\|^2.$$

Покажите также, что

$$4(x, y)=\|x+y\|^2-\|x-y\|^2.$$

Пусть $\{x_n\}$ есть последовательность точек в линейном пространстве со скалярным произведением. Тогда

(a) $\{x_n\}$ называется *последовательностью Коши*, если для любого $\varepsilon>0$ найдется такое $N=N(\varepsilon)$, что при всех n, $m\geqslant N$

$$\|x_n-x_m\|<\varepsilon,$$

(b) $\{x_n\}$ называется *сходящейся последовательностью*, если в этом пространстве существует такая точка x, что для

каждого $\varepsilon > 0$ найдется такое $N = N(\varepsilon)$, что при всех $n \geqslant N$

$$\| x - x_n \| < \varepsilon.$$

Упражнение 9. Покажите, что сходящаяся последовательность будет последовательностью Коши.

Чтобы показать, что обратное утверждение неверно, поступим следующим образом [1]).

Рассмотрим линейное пространство функций, непрерывных на интервале $[0, 1]$, определяемое естественными операциями сложения функций и умножения функции на число, и определим в нем скалярное произведение как интеграл от произведения функций на интервале $[0, 1]$. Как известно из анализа, пределом последовательности таких функций может быть и разрывная функция, что и показывает неверность обратного утверждения.

Пространство, для которого все последовательности Коши являются сходящимися, называется *полным*. Полное линейное пространство со скалярным произведением называется *гильбертовым пространством*.

До сих пор неявно рассматривались пространства, элементы которых являются точками на действительной оси, векторами или матрицами. Чтобы получить гильбертово пространство, удобное для метода конечных элементов, необходимо ввести такое пространство, в котором точки представляют собой функции. Наиболее употребительные функциональные пространства могут быть определены по аналогии с простейшим гильбертовым пространством, обозначаемым через $\mathscr{L}_2(R)$. Пусть для простоты R обозначает интервал (a, b) на действительной оси. Функция $f(x)$ является точкой этого пространства только тогда, когда интеграл

$$\int_a^b f^2(x)\, dx$$

конечен. Такая функция называется *измеримой* [2]). Для любых двух точек $u(x)$ и $v(x)$ скалярное произведение определяется как

$$(u, v) = \int_n^b u(x)\, v(x)\, dx,$$

---

[1]) В оригинале по этому поводу приводится другой пример, который оказался ошибочным. — *Прим. перев.*

[2]) Обычно такую функцию называют суммируемой с квадратом. — *Прим. перев.*

а норма как

$$\|u\|^2 = \int_a^b u^2(x)\,dx.$$

Сложение определяется как $(u+v)(x) = u(x) + v(x)$.

Упражнение 10. Покажите, что если $\|u\|$ и $\|v\|$ конечны, то скалярное произведение (u, v) также будет конечным.

Подмножество линейного пространства, которое само является линейным пространством, называется *подпространством.*

Упражнение 11. Пусть $\mathscr{L}_2(R)$ — пространство, определенное выше. Будут ли подпространствами следующие его подмножества:

(I) функций u, таких, что $u(a) = 0$.

(II) функций u, таких, что $(du/dx)_{x=a} = 0$ и $u(b) = 0$?

Упражнение 12. Пусть K есть подпространство линейного пространства $\mathscr{H}$, не совпадающее с ним самим, и пусть X — множество точек вида $\{x + h\}$, где x — некоторая фиксированная точка, принадлежащая $\mathscr{H}$, но не принадлежащая K, а h — любая точка из K. Покажите, что множество X, обозначаемое через $\{x\} \oplus K$, не является линейным пространством.

Множество X называется *линейным многообразием.*

Отображение T гильбертова пространства $\mathscr{H}$ на само себя называется линейным оператором, если

(I) $T(x+y) = T(x) + T(y)$ (для всех $x,\ y \in \mathscr{H}$),

(II) $T(\alpha x) = \alpha T(x)$ (для любого скаляра α).

Линейный оператор T называется *ограниченным*, если существует такая постоянная $M > 0$, что

$$\|Tx\| \leqslant M\|x\| \quad (\text{для всех } x \in \mathscr{H}),$$

и наименьшее из всех возможных M значение называется нормой оператора и обозначается через $\|T\|$; ясно, что

$$\|T\| = \sup_{x \neq 0} \left\{ \frac{\|Tx\|}{\|x\|} \right\} = \sup_{\substack{x \neq 0 \\ \|x\| \leqslant 1}} \left\{ \frac{\|Tx\|}{\|x\|} \right\} = \sup_{\|x\|=1} \{\|Tx\|\}.$$

Ограниченный линейный оператор *непрерывен*: это означает, что если точка x является пределом последовательности $\{x_n\}$, то точка Tx будет пределом последовательности $\{Tx_n\}$.

Пусть F — линейное отображение прямого произведения пространств $\mathscr{H}_1 \times \mathscr{H}_2$ в пространство $\mathscr{H}_3$, т. е. для любых $x_1 \in$

$\in \mathcal{H}_1$, $x_2 \in \mathcal{H}_2$ значение $F(x_1, x_2) \in \mathcal{H}_3$ и F линейно по каждому из аргументов; тогда F будет ограниченным, если существует такое $M > 0$, что

$$\| F(x_1, x_2) \|_{\mathcal{H}_3} \leqslant M \| x_1 \|_{\mathcal{H}_1} \| x_2 \|_{\mathcal{H}_2},$$

и $\|F\|$ есть наименьшее из таких M. Будем обозначать через $\mathcal{L}(\mathcal{H}_1; \mathcal{H}_2)$ пространство ограниченных линейных отображений из $\mathcal{H}_1$ в $\mathcal{H}_2$. Пространство $\mathcal{L}(\mathcal{H}; \mathbb{R})$ *ограниченных линейных функционалов* называется *двойственным* к $\mathcal{H}$ и обозначается через $\mathcal{H}'$. Элементы пространства $\mathcal{L}(\mathcal{H} \times \mathcal{H}; \mathbb{R})$ есть билинейные формы (заданные на $\mathcal{H}$).

Упражнение 13. Пусть $E \in \mathcal{L}(\mathcal{H}_1 \times \mathcal{H}_2; \mathcal{H}_3)$; определим $F \in \mathcal{L}(\mathcal{H}_1; \mathcal{H}_3)$, так, что при некотором фиксированном $v \in \mathcal{H}_2$

$$F(u) = E(u, v) \quad \text{(для всех } u \in \mathcal{H}_1\text{)}.$$

Докажите, что

$$\| F \| \leqslant \| E \| \| v \|_{\mathcal{H}_2}.$$

Отметим, что в дальнейшем для краткости нижние индексы у норм операторов часто будут опускаться.

Теорема 1.1. (Теорема Рисса о представлении функционала) (см. Иосида, 1967, с. 132). *$F \in \mathcal{H}'$ тогда и только тогда, когда существует единственный вектор $v \in \mathcal{H}$, такой, что для всех $u \in \mathcal{H}$*

$$F(u) = (u, v)$$

и

$$\| F \|_{\mathcal{H}'} = \| v \|_{\mathcal{H}}.$$

Теорема Рисса устанавливает взаимно однозначное соответствие между $\mathcal{H}$ и $\mathcal{H}'$; такое соответствие называется изоморфизмом, а два пространства, связанные таким соответствием, т. е. имеющие одинаковую структуру, называются изоморфными. Из определения нормы оператора следует, что

$$\| F \|_{\mathcal{H}'} = \sup_{\| u \| \neq 0} \left\{ \frac{| F(u) |}{\| u \|_{\mathcal{H}}} \right\},$$

а из теоремы 1.1 следует, что существует такой элемент $v \in \mathcal{H}$, для которого

$$F(u) = (u, v).$$

Поскольку оказалось возможным сопоставить каждому элементу из $\mathcal{H}'$ единственный элемент из $\mathcal{H}$, не должна вызы-

вать недоумение и запись вида

$$\|v\|_{\mathscr{H}'} = \sup_{\|u\| \neq 0} \left\{ \frac{|(u, v)|}{\|u\|} \right\}.$$

Линейный оператор T, который отображает все гильбертово пространство $\mathscr{H}$ на подпространство K, являющееся лишь частью $\mathscr{H}$, называется проекцией, если он отображает точки из K на самих себя, т. е. если

$$Ty = y \quad (\text{для всех } y \in K).$$

Упражнение 14. Какие из следующих линейных операторов будут проекциями?

(I) Оператор T, отображающий двумерный вектор $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ в вектор $\begin{pmatrix} \alpha \\ 0 \end{pmatrix}$.

(II) Оператор Q, отображающий пространство $\mathscr{L}_2(R)$, $R = [a, b]$, на пространство линейных функций путем линейной интерполяции по значениям в концах интервала.

(III) Оператор S, переводящий пространство матриц порядка 2×2 в диагональные матрицы по формуле

$$S\left(\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}\right) = \begin{bmatrix} \alpha + \delta & 0 \\ 0 & \beta + \gamma \end{bmatrix}.$$

(IV) Оператор S^2, где S определен в (III).

Проекция P называется *ортогональной*, если для всех x из пространства $\mathscr{H}$ и всех y из его подпространства K

$$(x - Px, y) = 0 \quad (\text{обозначается как } (x - Px) \perp y),$$

т. е. если разность $x - Px$ ортогональна всем y из подпространства K. Говорят, что эта разность принадлежит ортогональному дополнению K, которое обозначается через $K^{\perp}$.

Лемма 1.1. *Если $\mathscr{H}$ есть гильбертово пространство, а K — любое его подпространство, то $K^{\perp}$ также будет гильбертовым пространством.*

Лемма 1.2. *Если P есть ортогональная проекция гильбертова пространства $\mathscr{H}$ на некоторое подпространство K, то $I - P$ будет ортогональной проекцией $\mathscr{H}$ на $K^{\perp}$ и для любого $x \in \mathscr{H}$ существуют $u \in K$ и $v \in K^{\perp}$, такие, что $x = u + v$, т. е. $\mathscr{H} = K \oplus K^{\perp}$.*

Упражнение 15. Докажите лемму 1.1 и лемму 1.2.

Теорема 1.2. *Ортогональная проекция гильбертова пространства на подпространство единственна, а величина $\|x - Px\|$*

является минимальным расстоянием от x до подпространства.

Доказательство. Допустим, что ортогональная проекция P не единственна. Тогда существует другая ортогональная проекция Q, для которой

$$(x - Qx,\ y) = 0 \quad (\text{при всех } y \in K).$$

Так как $Px - Qx \subset K$, то

$$\begin{aligned} 0 &= (x - Qx,\ Px - Qx) = \\ &= (x - Px,\ Px - Qx) + (Px - Qx,\ Px - Qx) = \\ &= 0 + \| Px - Qx \|^2 > 0, \end{aligned}$$

поскольку $Px \neq Qx$, и получается противоречие. Следовательно, исходное предположение неверно и ортогональная проекция единственна.

Пусть теперь v есть любая точка в K, отличная от Px. Тогда, как и выше,

$$(v - Px) \perp (x - Px).$$

Следовательно,

$$\begin{aligned} \| x - v \|^2 &= \| (x - Px) + (Px - v) \|^2 = \\ &= \| x - Px \|^2 + 2(x - Px,\ Px - v) + \| Px - v \|^2 = \\ &= \| x - Px \|^2 + \| Px - v \|^2 > \| x - Px \|^2, \end{aligned}$$

а это и является нужным для нас результатом.

Следствие (I). $\|Px\|^2 = (x, Px)$ *и* $\|x - Px\|^2 = \|x\|^2 - (x, Px)$.

Следствие (II). *Для любого $x \in \mathcal{H}$ существуют единственные $u_0 \in K$ и $v_0 \in K^\perp$, такие, что $x = u_0 + v_0$ и*

$$\min_{u \in K} \| x - u \|^2 = \| x - u_0 \|^2 = \| v_0 \|^2. \tag{1.21}$$

В (1.21) можно поменять ролями K и $K^\perp$, и тогда получится другое минимизирующее соотношение

$$\min_{v \in K^\perp} \| x - v \| = \| u_0 \|^2.$$

Поскольку

$$\| x \|^2 = \| u_0 \|^2 + \| v_0 \|^2,$$

последняя задача минимизации может быть сведена к нахождению

$$\max_{v \in K^\perp} \{ \| x \|^2 - \| x - v \|^2 \}$$

или

$$\max_{v \in K^{\perp}} \{2(x, v) - \| v \|^2\}, \tag{1.22}$$

а эта задача имеет то же самое решение, что и задача (1.21).

Линейный оператор T называется *положительно определенным*, если

$$(Tx, x) > 0 \quad (\text{при всех } x \neq 0)^{1}),$$

и называется *положительно полуопределенным*, если

$$(Tx, x) \geqslant 0 \quad (\text{при всех } x \neq 0).$$

Например, отображение $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ в $\begin{pmatrix} \alpha/2 \\ \beta/2 \end{pmatrix}$ будет положительно определенным, тогда как отображение $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ в $\begin{pmatrix} \alpha \\ 0 \end{pmatrix}$ будет положительно полуопределенным, так как

$$(Tx, x) = \alpha^2$$

и обращается в нуль при $\alpha = 0$ и произвольном β.

Сопряженным к T называется такой оператор T^*, для которого

$$(Tx, y) = (x, T^* y) \quad (\text{при всех } x, y).$$

Если $(Tx, y) = (x, Ty)$, то T называется *самосопряженным* оператором.

1.3. Аппроксимирующие подпространства

Рассмотрим гильбертово пространство $\mathcal{H}$, и пусть $S = \{x_1, \ldots, x_N\}$ есть множество N элементов из $\mathcal{H}$. Эти элементы называются линейно независимыми, если равенство

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_N x_N = 0$$

имеет место только при нулевых значениях коэффициентов α_i $(i = 1, 2, \ldots, N)$. Если для каждого элемента $x \in \mathcal{H}$ существуют такие коэффициенты β_i, что

$$x = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_N x_N,$$

то говорят, что множество S образует базис в $\mathcal{H}$, а $\mathcal{H}$ есть N-мерное пространство.

---

[1]) Другие авторы используют в определении неравенство $(Tx, x) \geqslant \gamma \|x\|^2$ при $\gamma > 0$.

Например, пространство векторов вида $\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$ является двумерным пространством. Множество $S_1 = \left\{e_1 = \begin{pmatrix}0\\1\end{pmatrix},\ e_2 = \begin{pmatrix}1\\0\end{pmatrix}\right\}$ образует в нем базис, точно так же, как и множество $S_2 = \left\{f_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix},\ f_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}\right\}$. Оба этих базиса изображены на рис. 5.

Упражнение 16. Докажите, что в одном пространстве два различных базиса должны иметь одинаковое число элементов.

Важность конечномерных функциональных пространств становится понятной, если упомянутые в разделе 1.1 аппроксимирующие функции рассматривать как элементы пространства $\mathscr{L}_2(R)$. Например, если интервал $R = [a, b]$ разбит точками x_i $(i = 0, 1, \ldots, n)$, то можно построить множество эрмитовых функций, которые будут кусочно-линейными на этом интервале. Любое линейно независимое множество из $(n+1)$ таких функций образует базис в пространстве $H^{(1)}(\Pi, R)$. Нетрудно показать, что H является полным подпространством в $\mathscr{L}_2(R)$ (мы предлагаем читателю сделать это в качестве упражнения), и поэтому H является $(n+1)$-мерным подпространством в $\mathscr{L}_2(R)$. Пирамидальные функции (1.3) представляются естественным базисом в H для нахож-

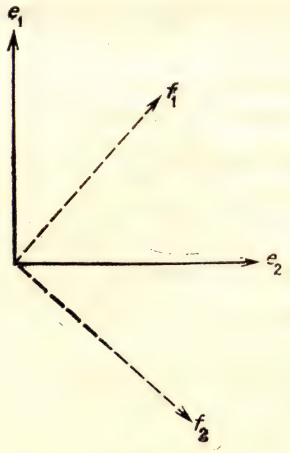
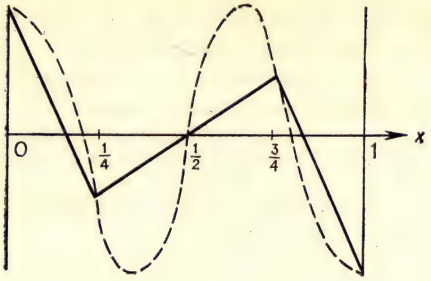


Рис. 5.



---- Cos $3\pi x$

—— $D_3(x)$

Рис. 6.

дения параметров

$$f_i \quad (i = 0, 1, \ldots, n).$$

Упражнение 17. Пусть разбиение Π задано точками $x = i/4$ $(i = 0, 1, \ldots, 4)$ и функции $D_i(x)$ определены как

$$D_i(x) = \sum_{k=0}^{4} \alpha_k^{(i)} \varphi_k \quad (i = 0, 1, \ldots, 4),$$

где если $\alpha^{(i)} = (\alpha_0^{(i)}, \ldots, \alpha_4^{(i)})$, то $\alpha^{(0)}(1, 1, 1, 1, 1)$, $\alpha^{(1)} = (1, 1/\sqrt{2}, 0, -1/\sqrt{2}, -1)$, $\alpha^{(2)} = (1, 0, -1, 0, 1)$, $\alpha^{(3)} = (1, -1/\sqrt{2}, 0, 1/\sqrt{2}, -1)$, $\alpha^{(4)} = (1, -1, 1, -1, 1)$.
На рис. 6 приведена функция $D_3(x)$ на интервале $[0, 1]$ вместе с функцией $\cos(3\pi x)$.

(I) Постройте приближенные графики $D_i(x)$ $(i = 0, 1, \ldots, 4)$.

(II) Вычислите скалярные произведения $(D_i(x), D_j(x))$ и затем рассмотрите $D_i(x)$ как базис в H и найдите коэффициенты g_i в разложении

$$g(x) = \sum_{i=0}^{4} g_i D_i(x).$$

Упражнение 18. Постройте базисы для пространства кубических сплайнов и пространства кусочно-линейных функций, заданных на треугольной сетке.

Все упомянутые в разделе 1.1 аппроксимирующие функции являются частными случаями общей задачи приближения. Эта задача состоит в том, что каждому элементу f гильбертова пространства $\mathcal{H}$ ставится в соответствие единственный элемент $\hat{f}$ аппроксимирующего N-мерного подпространства K

(например, если $K = H$, то $N = n + 1$). Элемент $\tilde{f}$ называется K-приближением f. Отображение f в $\tilde{f}$ обычно является линейным и при наличии линейности представляет собой проекцию. Например, если $K = H$, то $\tilde{f}$ можно получить путем линейной интерполяции между точками разбиения. Это единственное отображение, являющееся также и проекцией.

Упражнение 19. Постройте отображение из $\mathscr{L}_2(R)$ в H, которое не является проекцией.

Если аппроксимация любого элемента $f \in \mathscr{L}_2(R)$ уже построена, имеет смысл задать вопрос о том, насколько она хороша. Аппроксимация является *наилучшей*, если элемент $\tilde{f} \in K$ оказывается таким, что норма ошибки $f - \tilde{f}$ минимальна. В теореме 1.2 было показано, что если P есть ортогональная проекция на K, то $\|f - Pf\|$ является минимальным расстоянием от f до подпространства K, и поэтому $\tilde{f} = Pf$ представляет собой наилучшую аппроксимацию. Так как предполагается, что K имеет конечную размерность, то это предположение может быть использовано для построения наилучшей аппроксимации. Пусть $S = \{f_1, \ldots, f_N\}$ образует базис в K. Так как

$$(f - \tilde{f}, g) = 0$$

для всех $g \in K$, то

$$\left(f - \tilde{f}, \sum_k^N \beta_k f_k\right) = 0$$

для всех возможных последовательностей коэффициентов β_k $(k = 1, \ldots, N)$, и поэтому

$$(f - \tilde{f}, f_k) = 0 \quad (k = 1, \ldots, N). \tag{1.23}$$

Если $\tilde{f} = \sum_{i=1}^{N} \alpha_i f_i$, то

$$(f, f_k) - \sum_{i=1}^{N} \alpha_i (f_i, f_k) = 0 \quad (k = 1, \ldots, N).$$

Это может быть переписано как

$$G\alpha = \mathbf{b}, \tag{1.24}$$

где

$$G = [g_{ik}] = [(f_i, f_k)],$$

$$\alpha = \{\alpha_1, \ldots, \alpha_N\}^T$$

и

$$\mathbf{b} = \{(f, f_1), \ldots, (f, f_N)\}^T.$$

Матрица G называется матрицей Грама, а система (1.24) называется *нормальной.*

Упражнение 20. Покажите, что для $\mathscr{H} = \mathscr{L}_2(R)$ наилучшая аппроксимация $\tilde{f}$, определенная условиями (1.23), является также наилучшей в смысле метода наименьших квадратов. Отметим, что нормальную систему (1.24) не рекомендуется использовать для вычисления решения задачи по методу наименьших квадратов, так как с ростом ее порядка ее обусловленность быстро ухудшается.

Упражнение 21. Покажите, что можно определить гильбертово пространство $\mathscr{L}_2^w(R)$ с помощью скалярного произведения

$$(u, v)_w = \int_a^b w(x)\, u(x)\, v(x)\, dx,$$

где $w(x) \geqslant 0$ есть подходящая весовая функция. Далее получите нормальную систему, которая определяет наилучшую в смысле метода наименьших квадратов с весом аппроксимацию для функции $f(x)$ из $\mathscr{L}_2^w(R)$.

Упражнение 22. Покажите, что для измеримых функций, имеющих измеримую первую производную, можно определить гильбертово пространство $\mathscr{H}_2^{(1)}(R)$ с помощью скалярного произведения

$$(u, v)_1 = \int_a^b \{u(x)\, v(x) + u'(x)\, v'(x)\}\, dx$$

и нормы

$$\|u\|_1^2 = \int_a^b \{u^2(x) + u'^2(x)\}\, dx,$$

где штрих означает дифференцирование по x. Получите нормальную систему для наилучшей аппроксимации функции $f(x)$ из этого пространства. Пространство $\mathscr{H}_2^{(1)}(R)$ представляет собой пример *пространства Соболева.*

Методы аппроксимации, включающие решение нормальной системы с целью получения ортогональной проекции функции на конечномерное подпространство, называются проекционными.

ГЛАВА 2

ВАРИАЦИОННЫЕ ПРИНЦИПЫ

2.1. Введение

Вариационные принципы встречаются во многих физических и других задачах, и методы приближенного решения таких задач часто основаны на соответствующих вариационных принципах. Математически вариационный принцип состоит в том, что интеграл от некоторой функции имеет меньшее (или большее) значение для реального состояния системы, чем для любого возможного состояния, допускаемого основными условиями системы. Подынтегральная функция зависит от координат, амплитуд поля и их производных, а интегрирование осуществляется по области, покрываемой координатами системы, среди которых, возможно, есть и время. Задача определения минимума интеграла часто сводится к решению одного или нескольких дифференциальных уравнений с частными производными при соответствующих граничных условиях. Цель нашей книги не в том, чтобы рассматривать приближенные методы решения этих дифференциальных уравнений как способ решения исходных физических задач, сформулированных в виде вариационных принципов. Вместо этого мы намерены описать приближенные методы, которые основаны непосредственно на вариационных принципах.

В качестве примера таких экстремальных задач рассмотрим двойной интеграл

$$I(u) = \iint_R F(x, y, u, u_x, u_y)\, dx\, dy, \tag{2.1}$$

где u — непрерывная вместе со всеми производными до второго порядка функция, значения которой на границе области R заданы. Область R здесь — ограниченная область на плоскости (x, y). Сравнительно легко показать (см. Курант и Гильберт, 1951), что необходимое условие экстремума $I(u)$ состоит в том, что функция $u(x, y)$ должна удовлетворять уравнению Эйлера — Лагранжа

$$\frac{\partial}{\partial x} F_{u_x} + \frac{\partial}{\partial y} F_{u_y} - F_u = 0. \tag{2.2}$$

Из многих решений этого уравнения выбирается то, которое удовлетворяет заданным граничным условиям. К примеру, для

$F = \frac{1}{2}(u_x^2 + u_y^2)$ уравнение (2.2) сводится к уравнению Лапласа

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

Причина требования непрерывности вторых производных u теперь ясна; непрерывность обеспечивает существование уравнения Эйлера. Однако приближенные методы, основанные на минимизации $I(u)$ в виде (2.1), требуют только непрерывности u и кусочной непрерывности ее первых производных.

Вернемся теперь к основной задаче вариационного принципа: найти такую функцию из допустимого класса функций, что некоторый определенный интеграл по замкнутой области R, зависящий от функции и ее производных, принимает максимальное или минимальное значение. Это есть обобщение элементарной теории вычисления максимумов и минимумов, которая состоит в нахождении точки замкнутой области, в которой функция имеет максимальное или минимальное значение в некоторой окрестности в этой области. Определенный интеграл в вариационном принципе есть пример *функционала* и зависит от всего поведения функции в целом, а не от числа переменных. Область определения функционала есть пространство допустимых функций. Главная трудность вариационного подхода состоит в том, что задачи, которые могут быть естественно сформулированы как вариационные, могут не иметь решений. Математически это выражается незамкнутостью пространства допустимых функций. *Поэтому в вариационном принципе нельзя предполагать существование максимума или минимума.* В этой книге мы, однако, имеем дело с приближенными решениями вариационных задач. Они получаются при рассмотрении некоторого замкнутого подмножества пространства допустимых функций для получения верхней и нижней оценок точного решения вариационной задачи.

Одно очевидное преимущество вариационного подхода состоит в том, что нужно налагать менее жесткие требования на непрерывность решения. Этот парадокс разъясняется в разд. 1.3 книги Стренга и Фикса (1973) и в гл. 2 книги Клегга (1967). Полезным следствием более слабых требований непрерывности является то, что в вариационном подходе легче строить приближенные решения. Большая часть данной книги состоит в описании таких приближенных методов.

Оцениваемый в вариационном принципе интеграл берется по пространству, которое может иметь координату-время. Мы рассмотрим сначала вариационные задачи, не содержащие время. Эти вариационные задачи обычно выражают принцип минимума потенциальной энергии при нахождении состояния

устойчивого равновесия во многих классических задачах математической физики. В этой главе содержатся только те вопросы теории вариационных принципов, которые имеют отношение к главной теме книги. Никаких доказательств или подробных обсуждений здесь нет, и в случае особой заинтересованности читателю следует обратиться к соответствующим разделам книг Куранта и Гильберта (1951), Морса и Фешбаха (1958), Хилдебранда (1965), Шехтера (1971) и Клегга (1967).

2.2. Стационарные задачи

Дифференциальное уравнение, которое связано с вариационным принципом, известно как уравнение *Эйлера — Лагранжа*. Оно является необходимым, реже — достаточным условием, которому должна удовлетворять функция, максимизирующая или минимизирующая определенный интеграл. В простейшей задаче вариационного исчисления требуется найти минимум интеграла

$$I(u) = \int_{x_0}^{x_1} F(x, u(x), u'(x))\, dx,$$

где граничные значения $u(x_0)$ и $u(x_1)$ заданы, а штрих означает дифференцирование по x. Необходимым (но не достаточным) условием существования минимума является дифференциальное уравнение

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} = 0,$$

которому должна удовлетворять функция u, доставляющая минимум функционалу $I(u)$. Приведем обобщение этого факта для следующих ситуаций.

(1) *Случай двух неизвестных функций.* Минимизируемый интеграл есть

$$I(u, v) = \int_{x_0}^{x_1} F(x, u(x), v(x), u'(x), v'(x))\, dx,$$

где значения $u(x_0)$, $u(x_1)$, $v(x_0)$, $v(x_1)$ заданы. Необходимые условия таковы:

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} = 0,$$

$$\frac{\partial F}{\partial v} - \frac{d}{dx}\frac{\partial F}{\partial v'} = 0.$$

(2) *Случай функции двух переменных.* Минимизируется интеграл

$$I(u) = \iint_R F(x, y, u(x, y), u_x(x, y), u_y(x, y))\, dx\, dy,$$

где u принимает заданные значения на границе области интегрирования R. Необходимое условие минимума:

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\frac{\partial F}{\partial u_x} - \frac{\partial}{\partial y}\frac{\partial F}{\partial u_y} = 0.$$

(3) *Наличие высших производных.* В задачах со вторыми производными минимизируется интеграл

$$I(u) = \int_{x_0}^{x_1} F(x, u(x), u'(x), u''(x))\, dx,$$

где значения $u(x_0)$, $u'(x_0)$, $u(x_1)$, $u'(x_1)$ заданы, а соответствующее необходимое условие есть

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} + \frac{d^2}{dx^2}\frac{\partial F}{\partial u''} = 0.$$

(4) *Условный экстремум.* В таких вариационных задачах функция $u(x)$ должна удовлетворять ряду дополнительных условий. Здесь ищется экстремум одного интеграла при условии, что другие интегралы сохраняют постоянные значения. Такие задачи называют *изопериметрическими.* В качестве примера рассмотрим задачу о нахождении экстремума интеграла

$$I(u) = \int_{x_0}^{x_1} F(x, u(x), u'(x))\ dx$$

при условии, что $u(x)$ удовлетворяет уравнению

$$\int_{x_0}^{x_1} G(x, u(x), u'(x))\, dx = \alpha, \tag{2.3}$$

где α — заданная константа. Необходимое условие экстремума заключается в том, чтобы

$$\frac{\partial (F + \lambda G)}{\partial u} - \frac{d}{dx}\frac{\partial (F + \lambda G)}{\partial u'} = 0,$$

где численное значение параметра λ находится из условия (2.3). В качестве простого примера изопериметрической задачи приведем следующую. Необходимо найти форму провисающей однородной струны с закрепленными концами. Здесь

требуется найти кривую $u(x)$, проходящую через точки (x_0, u_0) и (x_1, u_1) и минимизирующую интеграл

$$\int_{x_0}^{x_1} u\,(1+u'^2)^{1/2}\,dx$$

при условии, что интеграл

$$\int_{x_0}^{x_1} (1+u'^2)^{1/2}\,dx$$

имеет фиксированное значение.

Упражнение 1. Покажите, что длина кривой, соединяющей две точки (x_0, u_0) и (x_1, u_1), есть

$$I(u)=\int_{x_0}^{x_1} (1+u'^2)^{1/2}\,dx.$$

Используя соответствующее уравнение Эйлера — Лагранжа, найдите путь наименьшей длины между этими точками.

Упражнение 2. Найдите кривую $u(x)$, проходящую через две точки (x_0, u_0) и (x_1, u_1) и дающую минимальную площадь поверхности вращения, при вращении кривой вокруг оси x.

Упражнение 3. Покажите, что уравнение

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2}+f(x,\,y,\,z)=0$$

является необходимым условием минимизации интеграла

$$I(u)=\iiint_R \frac{1}{2}\left(u_x^2+u_y^2+u_z^2-2uf(x,\,y,\,z)\right)dx\,dy\,dz,$$

когда на поверхности ∂R, содержащей объем R, функция $u(x, y, z)$ задана.

Прежде чем идти дальше, дадим несколько примеров вариационных принципов и эквивалентных им уравнений Эйлера — Лагранжа. В этих примерах функции определяются в области R с границей ∂R.

(1) *Задача Дирихле для уравнения Лапласа.*

$$I(u)=\iint_R \frac{1}{2}\left(u_x^2+u_y^2\right)dx\,dy,$$

$$u_{xx}+u_{yy}=0 \quad (u \text{ задана на } \partial R).$$

(2) *Нагруженная пластина с заделанным краем.* (Бигармонический оператор)

$$I(u)=\iint\limits_R \frac{1}{2}\left[u_{xx}^2+2u_{xy}^2+u_{yy}^2-2qu\right]dx\,dy$$

$$\left(u=\frac{du}{dn}=0 \text{ на } \partial R\right),$$

$$u_{xxxx}+2u_{xxyy}+u_{yyyy}=q(x,\ y).$$

Здесь $q(x, y)$ — нагрузка по нормали на пластину.

(3) *Теория упругости при малых деформациях.* (Плоское напряженное состояние.)

$$I(u,\ v)=\iint\limits_R \frac{1}{2}\Big[(1-\nu)(u_x^2+v_y^2)+\nu(u_x+v_y)^2+$$

$$+\frac{1}{2}(1-\nu)(u_y+v_x)^2\Big]dx\,dy,$$

$$u_{xx}+\nu v_{xy}+\frac{1}{2}(1-\nu)(u_{yy}+v_{xy})=0,\ (u,\ v \text{ заданы на } \partial R),$$

$$v_{yy}+\nu u_{xy}+\frac{1}{2}(1-\nu)(u_{xy}+v_{xx})=0.$$

(4) *Радиация* (e^u) *и диффузия молекул* (u^2).

$$I(u)=\iint\limits_R \frac{1}{2}\left(u_x^2+u_y^2+\begin{cases}2e^u\\ \frac{2}{3}u^3\end{cases}\right)dx\,dy,$$

$$u_{xx}+u_{yy}=\begin{cases}e^u\\ u^2\end{cases}\quad (u \text{ задана на } \partial R).$$

(5) *Задача Плато.* (Найти поверхность минимальной площади, ограниченную замкнутой кривой в трехмерном пространстве.)

$$I(u)=\iint\limits_R \frac{1}{2}(1+u_x^2+u_y^2)^{1/2}\,dx\,dy,$$

$$\nabla(\gamma_1)\nabla(u)=0,\ \gamma_1=(1+u_x^2+u_y^2)^{-1/2}\quad (u \text{ задана на } \partial R).$$

(6) *Неньютоновские жидкости.*

$$I(u)=\iint\limits_R \left[\frac{1}{2}(u_x^2+u_y^2)^{1+s}+cu\right]dx\,dy$$

$$\nabla(\gamma_2)\nabla u=c,\ \gamma_2=(u_x^2+u_y^2)^s,\ \left(-\frac{1}{2}\leqslant s\leqslant 0\right),\ (c-\text{const}),$$

$$(u \text{ задана на } \partial R),$$

(7) *Течение сжимаемой жидкости.*

$$I(p)=\iint_R p\,d\mathbf{x} \qquad \begin{cases} \rho \text{ — плотность,} \\ p \text{ — давление,} \\ \varphi \text{ — потенциал скорости.} \end{cases}$$

$$\nabla(\rho\nabla\varphi)=0$$

Из этих задач первые три — линейные, четвертая — слабо нелинейная, последние три — нелинейные.

2.3. Граничные условия

Стационарные задачи, описанные в предыдущем разделе, таковы, что искомая функция задана на границе и не может там варьироваться [1]). Однако во многих задачах функция не задана на границе, и применяют другие граничные условия. Рассмотрим, например, (Курант и Гильберт, 1951, стр. 163—164) вариационную задачу, состоящую в минимизации интеграла

$$I(u)=\int_{x_0}^{x_1} F(x,\ u,\ u')\,dx, \tag{2.4}$$

где u не задана в точках $x=x_0, x_1$. Необходимое условие минимизации $I(u)$ состоит в том, что $u(x)$ удовлетворяет уравнению Эйлера — Лагранжа

$$\frac{\partial F}{\partial u}-\frac{d}{dx}\frac{\partial F}{\partial u'}=0$$

и граничным условиям

$$\frac{\partial F}{\partial u'}=0 \quad (x=x_0,\ x_1).$$

Последние известны как *естественные* граничные условия, поскольку они следуют непосредственно из минимизации основного интеграла. Если граничные условия задачи не заданы непосредственно и не определены естественные граничные условия, минимизируемый функционал необходимо должным образом изменить.

Рассмотрим следующий модифицированный вид (2.4):

$$I(u)=\int_{x_0}^{x_1} F(x,\ u,\ u')\,dx+[g_1(x,\ u)]_{x=x_1}-[g_0(x,\ u)]_{x=x_0},$$

---

[1]) Часто такие граничные условия называют *главными*, в отличие от естественных, см. ниже. — *Прим. перев.*

где $g_0(x, u)$ и $g_1(x, u)$ — произвольные функции. Необходимым условием минимизации этого функционала является (см. Шехтер 1971, с. 35) уравнение

$$\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u'} = 0$$

с граничными условиями

$$\frac{\partial F}{\partial u'} + \frac{\partial g_0}{\partial u} = 0 \quad (x = x_0),$$

$$\frac{\partial F}{\partial u'} + \frac{\partial g_1}{\partial u} = 0 \quad (x = x_1).$$

Таким образом, с помощью функций $g_0(x, u)$ и $g_1(x, u)$ можно получить подходящие граничные условия задачи. Например, вариационная задача, эквивалентная дифференциальному уравнению

$$u'' + f(x) = 0$$

с граничными условиями

$$-u' + \alpha u = 0 \quad (x = x_0),$$

$$u' + \beta u = 0 \quad (x = x_1)$$

имеет функционал

$$I(u) = \int_{x_0}^{x_1} \left[\frac{1}{2}u'^2 - f(x)\,u\right] dx + \left[\frac{1}{2}\beta u^2\right]_{x=x_1} - \left[\frac{1}{2}\alpha u^2\right]_{x=x_0}.$$

Если теперь мы рассмотрим двумерную вариационную задачу, состоящую в минимизации интеграла

$$I(u) = \iint_R F(x, y, u, u_x, u_y)\, dx\, dy,$$

где u не задана на границе области R, необходимое условие минимума состоит в том, что u удовлетворяет дифференциальному уравнению

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\frac{\partial F}{\partial u_x} - \frac{\partial}{\partial y}\frac{\partial F}{\partial u_y} = 0$$

с естественными граничными условиями

$$\frac{\partial F}{\partial u_x}\frac{dy}{d\sigma} - \frac{\partial F}{\partial u}\frac{dx}{d\sigma} = 0$$

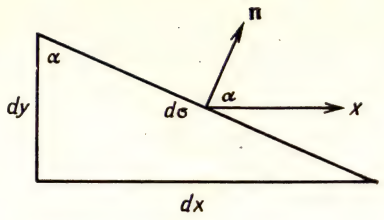


Рис. 7.

на кривой ∂R, являющейся границей R. Если угол между нормалью к ∂R и осью x равен α (см рис. 7), то $\cos\alpha = dy/d\sigma$ и $\sin\alpha = -dx/d\sigma$, где σ означает длину дуги вдоль границы. В более общем случае двух искомых функций u и v (Хильдебранд, 1965, стр. 135) получается дополнительное уравнение Эйлера — Лагранжа для v

$$\frac{\partial F}{\partial v} - \frac{\partial}{\partial x}\frac{\partial F}{\partial v_x} - \frac{\partial}{\partial y}\frac{\partial F}{\partial v_y} = 0$$

и дополнительное естественное граничное условие

$$\frac{\partial F}{\partial v_x}\frac{dy}{d\sigma} - \frac{\partial F}{\partial v_y}\frac{dx}{d\sigma} = 0.$$

Если граничные условия не определяют явно граничные значения u (или v) или не являются естественными, то снова необходимо изменять функционал. Рассмотрим это на примере двумерной задачи со вторыми производными в подынтегральном выражении. Пусть требуется найти функцию $u(x, y)$, доставляющую стационарное значение функционалу

$$I(u) = \iint_R F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy})\,dx\,dy + \int_{\partial R} G(x, y, u, u_\sigma, u_{\sigma\sigma}, u_n)\,d\sigma,$$

где $\partial/\partial\sigma$ и $\partial/\partial n$ — операторы дифференцирования по направлениям касательной и нормали к кривой ∂R. Интегрирование по ∂R осуществляется так, что область R при движении вдоль ∂R оказывается слева (движение по ∂R против часовой стрелки). Необходимое условие минимума $I(u)$ — уравнение Эйлера — Лагранжа

$$F_u - \frac{\partial}{\partial x}F_{u_x} - \frac{\partial}{\partial y}F_{u_y} + \frac{\partial^2}{\partial x^2}F_{u_{xx}} + \frac{\partial^2}{\partial x\,\partial y}F_{u_{xy}} + \frac{\partial^2}{\partial y^2}F_{u_{yy}} = 0$$

с граничными условиями

$$\left[F_{u_x}-\frac{\partial}{\partial x}F_{u_{xx}}\right]\frac{dy}{d\sigma}-\left[F_{u_y}-\frac{\partial}{\partial y}F_{u_{yy}}\right]\frac{dx}{d\sigma}-$$
$$-\left\{\frac{\partial}{\partial\sigma}\left(F_{u_{xx}}-F_{u_{yy}}\right)\right\}\frac{dx}{d\sigma}\frac{dy}{d\sigma}+$$
$$+\frac{1}{2}\left\{\frac{\partial}{\partial\sigma}F_{u_{xy}}\left\{\left(\frac{dx}{d\sigma}\right)^2-\left(\frac{dy}{d\sigma}\right)^2\right\}\right\}+$$
$$+\frac{1}{2}\left\{\left(\frac{\partial}{\partial x}F_{u_{xy}}\right)\frac{dx}{d\sigma}-\left(\frac{\partial}{\partial y}F_{u_{xy}}\right)\frac{dy}{d\sigma}\right\}+$$
$$+G_u-\frac{\partial}{\partial\sigma}G_{u_\sigma}+\frac{\partial^2}{\partial\sigma^2}G_{u_{\sigma\sigma}}=0, \quad (2.5)$$

$$\frac{\partial G}{\partial u_n}+\frac{\partial F}{\partial u_{xx}}\left(\frac{dy}{d\sigma}\right)^2+\frac{\partial F}{\partial u_{yy}}\left(\frac{dx}{d\sigma}\right)^2+\frac{\partial F}{\partial u_{xy}}\frac{dx}{d\sigma}\frac{dy}{d\sigma}=0. \quad (2.6)$$

Функция G выбирается так, чтобы граничные условия (2.5) и (2.6) соответствовали естественным граничным условиям задачи.

2.4. Смешанные вариационные принципы [1])

Рассмотрим теперь теорию упругости при малых деформациях и принцип минимума потенциальной энергии, изложенный в примере (2) разд. 2.6. Этот принцип предполагает, что имеют место определенные соотношения между напряжением и деформацией и между деформациями и перемещениями, а также выполнены кинематические граничные условия.

Если ослабить два последних условия и рассматривать их как множество ограничений, то можно написать модифицированный функционал вида

$$I(u,\ v,\ w)=I_p-\iiint_R\left(\varepsilon_x-\frac{\partial u}{\partial x}\right)\alpha_1\,d\mathbf{x}-\ldots-$$
$$-\iiint_R\left(\gamma_{zx}-\frac{\partial u}{\partial z}-\frac{\partial w}{\partial x}\right)\alpha_6\,d\mathbf{x}-$$
$$-\int_{S_d}(u-\bar u)\beta_1\,d\sigma-\int_{S_d}(v-\bar v)\beta_2\,d\sigma-\int_{S_d}(w-\bar w)\beta_3\,d\sigma, \quad (2.7)$$

где I_p — функционал потенциальной энергии, $\alpha_1, \ldots, \alpha_6$, β_1, β_2, β_3 — множители Лагранжа. Независимыми варьируемыми величинами являются три перемещения, шесть деформаций и девять множителей Лагранжа. Из (2.7) можно получить много смешанных вариационных принципов. В частности, если

$$\alpha_1=\sigma_x,\ \ldots,\ \alpha_6=\tau_{zx},\quad \beta_1=\sigma_x,\quad \beta_2=\sigma_y,\quad \beta_3=\sigma_z,$$

[1]) Этот раздел лучше читать после раздела 2.6. — *Прим. перев.*

то получается принцип Ху — Васидзу, а при

$$\alpha_1 = \sigma_x, \ \ldots, \ \alpha_6 = \tau_{zx}$$

возникает принцип Рейснера — Хелингера.

Функционал, связанный с принципом Рейснера — Хелингера, можно записать как

$$I_{RH} = I_G - \iiint_R \left(\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} + X\right) u \, d\mathbf{x} -$$

$$- \iiint_R \left(\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} + Y\right) v \, d\mathbf{x} -$$

$$- \iiint_R \left(\frac{\partial \tau_{zx}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \sigma_z}{\partial z} + Z\right) w \, d\mathbf{x} +$$

$$+ \int_{S_\sigma} (l\sigma_x + m\tau_{xy} + n\tau_{zx} - \bar{X}) u \, d\sigma +$$

$$+ \int_{S_\sigma} (l\tau_{xy} + m\sigma_y + n\tau_{yz} - \bar{Y}) v \, d\sigma +$$

$$+ \int_{S_\sigma} (l\tau_{zx} + m\tau_{yz} + n\sigma_z - \bar{Z}) w \, d\sigma, \qquad (2.8)$$

где I_c — функционал дополнительной энергии.

Принципы Ху — Васидзу и Рейснера — Хелингера являются смешанными принципами и утверждают стационарность, а не экстремальность значений функционала для реальных состояний. Несмотря на это, в приближенных методах (например, в методе конечных элементов), основанных на смешанных вариационных принципах, достигается примерно одинаковая точность таких величин, как перемещения и напряжения, тогда как при использовании принципа минимума энергии хорошая точность может быть получена либо для перемещений, либо для напряжений, но не для обоих одновременно.

Более полное обсуждение смешанных вариационных принципов и полное определение величин, использованных в (2.7) и (2.8), имеется в книге Васидзу (1968) и Табаррока (1973).

2.5. Вариационные принципы в нестационарных задачах

Наиболее важным и фундаментальным из таких вариационных принципов, использующих и время в качестве независимой переменной, является принцип *Гамильтона*, из которого могут быть выведены основные уравнения для большого

числа физических явлений. Принцип Гамильтона утверждает, что реальное движение системы материальных точек с момента времени t_0 до момента t_1 таково, что интеграл по времени от разности кинетической и потенциальной энергий системы стационарен для траектории этого движения. Математически это означает, что интеграл от лагранжиана L

$$I = \int_{t_0}^{t_1} L\,dt = \int_{t_0}^{t_1} (T - V)\,dt,$$

где T и V есть соответственно кинетическая и потенциальная энергии системы, имеет стационарное значение для реальной траектории по сравнению с близкими возможными траекториями. Для системы с n обобщенными координатами $q_1, q_2, \ldots, q_n$ связанные с этим принципом уравнения Эйлера — Лагранжа оказываются такими:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_r}\right) - \frac{\partial}{\partial q_r}(T - V) = 0 \quad (r = 1, 2, \ldots, n).$$

На них обычно ссылаются как на уравнения Лагранжа движения системы.

В качестве простого примера применения этой теории к сплошной среде рассмотрим случай гибкой струны, на которую действует постоянное натяжение τ. Концы струны закреплены, и она совершает малые колебания около положения устойчивого равновесия — интервала $0 \leqslant x \leqslant 1$ оси x. Если $u(x, t)$ — перемещение точки струны перпендикулярно оси x, то

$$T = \frac{1}{2}\rho \int_0^1 \left(\frac{\partial u}{\partial t}\right)^2 dx, \quad V = \frac{1}{2}\rho \int_0^1 c^2 \left(\frac{\partial u}{\partial x}\right)^2 dx,$$

где ρ — плотность струны и $c^2 = \tau/\rho$. Уравнение Эйлера — Лагранжа

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

является волновым уравнением. Таким образом, для струны волновое уравнение эквивалентно требованию минимальности усредненной разности полной кинетической и потенциальной энергий струны в предположении выполнения начальных и граничных условий задачи. Другими примерами применения этой теории служат колебания стержня, мембраны и пластины (Курант и Гильберт, 1951, стр. 215—221).

Вариационные принципы, включающие время как переменную, пока еще редко используются для численного реше-

ния эволюционных задач. Они обычно решаются полудискретными методами Галеркина (см. гл. 6).

Теперь перейдем к нестационарным *диссипативным* системам и покажем, как можно построить для них вариационные принципы. Применяемый метод состоит во введении сопряженной системы с отрицательным трением. Энергия, теряемая диссипативной системой, передается сопряженной системе, и поэтому полная энергия обеих систем сохраняется. Другой подход, основанный на ограниченных вариационных принципах, можно найти у Розена (1954). В качестве примера рассмотрим одномерный осциллятор с трением. Уравнение его движения есть

$$\ddot{x} + k\dot{x} + n^2 x = 0 \quad (k > 0).$$

Требуется найти вариационный принцип, для которого уравнение Эйлера — Лагранжа совпадало бы с этим уравнением. Это невозможно сделать, но если мы рассмотрим также сопряженный осциллятор (описываемый координатой x^*) с отрицательным трением и с уравнением движения

$$\ddot{x}^* - k\dot{x}^* + n^2 x^* = 0,$$

то для построенного чисто формально лагранжиана

$$L = \dot{x}\dot{x}^* - \frac{1}{2} k (x^*\dot{x} - x\dot{x}^*) - n^2 x x^*$$

уравнения Эйлера — Лагранжа совпадут с приведенными выше уравнениями движения обоих осцилляторов.

Другим важным примером диссипативной системы является задача о диффузии тепла. В одномерном случае она определяется уравнением

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.$$

Тогда сопряженную задачу определит уравнение

$$-\frac{\partial u^*}{\partial t} = \frac{\partial^2 u^*}{\partial x^2}.$$

Эти два уравнения являются уравнениями Эйлера — Лагранжа для формального лагранжиана

$$L = -\frac{\partial u}{\partial x}\frac{\partial u^*}{\partial x} - \frac{1}{2}\left(u^*\frac{\partial u}{\partial t} - u\frac{\partial u^*}{\partial t}\right).$$

2.6. Двойственные вариационные принципы

До сих пор наши вариационные принципы были «односторонними», т. е. приближенное решение всегда давало оценку теоретического решения задачи либо сверху, либо снизу. Однако часто можно для задачи построить два вариационных

принципа так, что некоторая величина d дает минимум для одного и максимум для другого вариационного принципа. Если d^u и d^l — приближенные решения для принципов минимальности и максимальности соответственно, то

$$d^l \leqslant d \leqslant d^u,$$

и поэтому мы имеем практический метод оценки величины d. Часто оказывается, что величина d имеет физический смысл.

Теперь мы дадим несколько примеров задач, в которых можно построить двойственные вариационные принципы.

(1) *Классическая задача Дирихле.* Здесь минимизируется функционал

$$I(u) = \iint_R \frac{1}{2}\left(u_x^2 + u_y^2\right) dx\, dy$$

на множестве непрерывных функций $u(x, y)$, имеющих в области R кусочно-непрерывные первые производные и заданные значения $u = f(\sigma)$ на ∂R, где σ — длина по ∂R, границе R. Двойственная задача состоит в максимизации функционала

$$J(v) = -\iint_R \frac{1}{2}\left(v_x^2 + v_y^2\right) dx\, dy - \int_{\partial R} v f'(\sigma)\, d\sigma$$

относительно непрерывных функций $v(x, y)$ с кусочно-непрерывными в R первыми производными, которые удовлетворяют естественным граничным условиям на ∂R. В этом примере

$$\min_u I(u) = \max_v J(v) = d.$$

Численная реализация этого двойственного принципа приводится в разд. 7.4 (C).

Упражнение 4. Поток несжимаемой невязкой жидкости параллелен оси x. Вычислите $I(u)$ и $J(v)$ в этой задаче, считая R квадратом $0 \leqslant x,\ y \leqslant 1$, и покажите, что их экстремальные значения совпадают. (Отметим, что u есть функция тока, а v — потенциал течения.)

Упражнение 5. Покажите, что необходимыми условиями максимума $J(v)$ являются уравнение Эйлера — Лагранжа

$$v_{xx} + v_{yy} = 0$$

и естественные граничные условия

$$v_y \frac{dx}{d\sigma} - v_x \frac{dy}{d\sigma} = f'(\sigma).$$

(2) *Теория упругости при малых деформациях* (Васидзу 1968). Пусть изотропное тело в трехмерном пространстве занимает область R, ограниченную замкнутой поверхностью ∂R. Компоненты объемных сил (на единицу объема) обозначим через X, Y, Z. Поверхность тела разбита на две части: S_σ, на которой в качестве граничных условий заданы внешние силы (на единицу площади) $(\bar{X}, \bar{Y}, \bar{Z})$, и S_d, на которой заданы перемещения $(\bar{u}, \bar{v}, \bar{w})$; при этом $\partial R = S_\sigma + S_d$. Тогда общая потенциальная энергия дается формулой

$$I_p = \iiint_R W(\varepsilon_x, \varepsilon_y, \varepsilon_z, \gamma_{yz}, \gamma_{zx}, \gamma_{xy})\, d\mathbf{x} - \\ - \iiint_R (Xu + Yv + Zw)\, dx - \int_{S_\sigma} (\bar{X}u + \bar{Y}v + \bar{Z}w)\, d\sigma,$$

где

$$W = \frac{E\nu}{2(1+\nu)(1-2\nu)}(u_x + v_y + w_z)^2 + \frac{E}{2(1+\nu)}(u_x^2 + v_y^2 + w_z^2) + \\ + \frac{E}{4(1+\nu)}[(v_z + w_y)^2 + (w_x + u_z)^2 + (u_y + v_x)^2];$$

а E — модуль Юнга и ν — коэффициент Пуассона для данной среды. Если объемные и поверхностные силы не изменяются в процессе вариаций, I_p достигает минимума при реальных перемещениях. В этом состоит принцип *минимума потенциальной энергии.*

Дополнительная энергия имеет вид

$$I_c = \iiint_R \varphi(\sigma_x, \sigma_y, \sigma_z, \tau_{yz}, \tau_{zx}, \tau_{xy})\, d\mathbf{x} - \int_{S_d} (X\bar{u} + Y\bar{v} + Z\bar{w})\, d\sigma,$$

где

$$\varphi = \frac{1}{2E}[(\sigma_x + \sigma_y + \sigma_z)^2 + 2(1+\nu)(\tau_{yz}^2 + \tau_{zx}^2 + \tau_{xy}^2 - \\ - \sigma_y\sigma_z - \sigma_z\sigma_x - \sigma_x\sigma_y)].$$

Если поверхностные перемещения остаются неизменными при вариациях, I_c достигает минимум при реальных напряжениях. Это — принцип *минимума дополнительной энергии.* С помощью этих двух принципов удобно оценивать коэффициент прямого влияния или обобщенное перемещение (Пиан, 1970).

Упражнение 6. Покажите, что $W = \varphi$, если выполнены следующие линейные соотношения между напряжением и де-

формацией:

$$E\varepsilon_x = \sigma_x - \nu(\sigma_y + \sigma_z), \quad \tau_{yz} = \frac{E}{2(1+\nu)}\gamma_{yz},$$

$$E\varepsilon_y = \sigma_y - \nu(\sigma_x + \sigma_z), \quad \tau_{zx} = \frac{E}{2(1+\nu)}\gamma_{zx},$$

$$E\varepsilon_z = \sigma_z - \nu(\sigma_x + \sigma_y), \quad \tau_{xy} = \frac{E}{2(1+\nu)}\gamma_{xy}.$$

Упражнение 7. Покажите, что необходимыми условиями минимума потенциальной энергии

$$I_p = \iint_R \Big[\frac{E\nu}{2(1+\nu)(1-2\nu)}(u_x + v_y)^2 + \\ + \frac{E}{2(1+\nu)}(u_x^2 + v_y^2) + \frac{E}{4(1+\nu)}(u_y + v_x)^2 \Big] dx\, dy$$

являются уравнения Эйлера — Лагранжа

$$(2-2\nu)u_{xx} + (1-2\nu)u_{yy} + v_{xy} = 0,$$
$$(2-2\nu)v_{yy} + (1-2\nu)v_{xx} + u_{xy} = 0$$

и граничные условия

$$2(1-\nu)u_x \cos\alpha + (1-2\nu)u_y \sin\alpha + (1-2\nu)v_x \sin\alpha + \\ + 2\nu v_y \cos\alpha = 0,$$

$$2\nu u_x \sin\alpha + (1-2\nu)u_y \cos\alpha + (1-2\nu)v_x \cos\alpha + \\ + 2(1-\nu)v_y \sin\alpha = 0$$

на границе ∂R области R (см. рис. 7).

(3) *Течение сжимаемой жидкости* (Севелл, 1969). Соответствующие объемные подынтегральные выражения, которые появляются в двойственных вариационных принципах, суть давление p и величина $p + \rho v^2$, где ρ — плотность, а v — скорость жидкости. Здесь

$$p = p(v_i, h, \eta),$$

где v_i $(i = 1, 2, 3)$ — составляющие скорости, h и η полная энергия и энтропия, приходящиеся на единицу массы соответственно. Из этих принципов следует, что

$$\frac{\partial p}{\partial v_i} = -Q_i \,(i = 1, 2, 3), \quad \frac{\partial p}{\partial h} = \rho, \quad \frac{\partial p}{\partial \eta} = -\rho T.$$

Здесь $Q_i = \rho v_i$ $(i = 1, 2, 3)$, а T — температура. Вводится функция

$$P = P(Q_i, h, \eta) = \sum_{i=1}^{3} Q_i v_i + p,$$

для которой

$$\frac{\partial P}{\partial Q_i} = v_i \ (i = 1, 2, 3), \quad \frac{\partial P}{\partial h} = \rho, \quad \frac{\partial P}{\partial \eta} = -\rho T.$$

Двойственные вариационные принципы, включающие p и P соответственно, могут усилить принципы экстремальности частных случаев течения сжимаемой жидкости.

Две попытки унификации двойственных принципов предприняты Севеллом (1969) и Артурсом (1970). Первый из них использует преобразования Лежандра (или инволютивные преобразования), второй — *каноническую теорию* уравнений Эйлера — Лагранжа. Превосходный обзор двойственных вариационных принципов вообще содержится в статье Нобла и Севелла (1972).

ГЛАВА 3

МЕТОДЫ АППРОКСИМАЦИИ

Вариационная формулировка вместе с присущими ей более слабыми требованиями непрерывности естественно переносится на приближенные методы решения, называемые обычно *прямыми методами* (Курант и Гильберт, 1951, стр. 154; Нечас, 1967). Применение этих методов сводит задачу к нахождению стационарных точек функции конечного числа вещественных переменных.

В этой главе, однако, мы рассматриваем лишь одно семейство прямых методов, а именно методы конечных элементов. Дается описание различных классов методов конечных элементов совместно с немногочисленными примерами из обширной области их применения; в гл. 7 приводится дополнительно еще несколько примеров. Вопросы точности и сходимости методов обсуждаются в гл. 5.

3.1. Метод Ритца

Создателем классического прямого метода обычно считают швейцарского математика В. Ритца (1878—1909). Если требуется решить вариационную задачу

$$\delta I(v) = 0 \quad (v(\mathbf{x}) \in \mathscr{H}), \tag{3.1}$$

где $\mathbf{x} = (x_1, \ldots, x_m)$ и $\mathscr{H}$ — пространство допустимых функций, приближенное решение u может быть получено, если ограничиться функциями из некоторого N-мерного подпространства $K_N \subset \mathscr{H}$. Сразу ясно, что если стационарная точка в (3.1) дает экстремум, то мы получаем оценку этого экстремального значения. Другими словами, если

$$\delta I(v_0) = 0,$$

тогда и только тогда, когда

$$I(v_0) = \min_{v \in \mathscr{H}} I(v) = d,$$

то

$$I(V) \geqslant d$$

для всех $V \in K_N$.

Если функции $\varphi_i(\mathbf{x})$ $(i = 1, \ldots, N)$ образуют базис подпространства K_N, то будем искать приближенное решение в виде

$$U(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \varphi_i(\mathbf{x}),$$

причем значения искомых параметров должны быть такими, что $I(U)$ стационарно относительно этих параметров α_i $(i = 1, 2, \ldots, N)$. Таким образом, получается система уравнений

$$\frac{\partial}{\partial \alpha_i} I\left(\sum_{j=1}^{N} \alpha_j \varphi_j\right) = 0 \quad (i = 1, \ldots, N).$$

Применим, например, метод Ритца к решению следующей задачи. В области $-\frac{1}{2}\pi < x,\ y < \frac{1}{2}\pi$ решить уравнение

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - 2 = 0 \tag{3.2}$$

при условии, что

$$\begin{aligned} u\left(x, \pm\tfrac{\pi}{2}\right) &= 0 \quad \left(|x| \leqslant \tfrac{\pi}{2}\right), \\ u\left(\pm\tfrac{\pi}{2}, y\right) &= 0 \quad \left(|y| \leqslant \tfrac{\pi}{2}\right). \end{aligned} \tag{3.2a}$$

Построим кусочно-билинейное приближенное решение по методу конечных элементов, используя метод Ритца, примененный к соответствующему этой задаче вариационному принципу

$$\delta I(v) = 0,$$

причем

$$I(v) = \iint_R \left[\tfrac{1}{2}(v_x^2 + v_y^2) + 2v\right] dx\, dy, \tag{3.3}$$

где

$$R = \left(-\tfrac{\pi}{2}, \tfrac{\pi}{2}\right) \times \left(-\tfrac{\pi}{2}, \tfrac{\pi}{2}\right).$$

В предыдущей главе было указано, что если на границе значения решения заданы (граничные условия Дирихле), то эти условия налагаются на пространство $\mathcal{H}$, тогда как в случае задания естественных граничных условий это не является обязательным. В данной задаче необходимо ограничиться пространством $\mathcal{H}$ функций, удовлетворяющих граничному условию $v = 0$.

Аппроксимирующие функции определены на области R, разбитой на $(T+1)^2$ квадратных элементов посредством $2T$ равно расположенных внутренних линий сетки, параллельных

осям (см. рис. 3). Тогда $N(=T^2)$ базисных функций $\varphi_{ij}(x, y)$ $(i, j = 1, \ldots, T)$ подпространства K_N, определенных в гл. 1 (упражнение 6), принадлежат, очевидно, пространству $\mathscr{H}$, так как они все обращаются в нуль на границе.

Приближенное решение имеет тогда вид

$$U(x, y) = \sum_{i, j=1}^{T} U_{ij}\varphi_{ij}(x, y), \tag{3.4}$$

где U_{ij} — значение приближенного решения в точке (x_i, y_j). Условием стационарности является система уравнений

$$\frac{\partial}{\partial U_{ij}} I\left(\sum_{k, l=1}^{T} U_{kl}\varphi_{kl}(x, y)\right) = 0 \quad (i, j = 1, \ldots, T). \tag{3.5}$$

Отсюда и из (3.3) получаем

$$\sum_{k, l=1}^{T} U_{kl} \iint_R \left\{\left(\frac{\partial\varphi_{kl}}{\partial x}\right)\left(\frac{\partial\varphi_{ij}}{\partial x}\right) + \left(\frac{\partial\varphi_{kl}}{\partial y}\right)\left(\frac{\partial\varphi_{ij}}{\partial y}\right)\right\} dx\,dy + \\ + 2\iint_R \varphi_{ij}\,dx\,dy = 0.$$

Непосредственное вычисление интегралов приводит к уравнениям

$$3U_{ij} - \frac{1}{3}\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1} U_{kl} + 2h^2 = 0 \quad (i, j = 1, \ldots, T),$$

где $U_{kl} = 0$, если $k, l = 0, T+1$. Эти уравнения можно записать через разностные операторы так:

$$\{\delta_x^2 I_y + \delta_y^2 I_x\} U_{ij} - 2h^2 = 0 \quad (i, j = 1, \ldots, T),$$

где δ_x^2, δ_x^2 — центрированные разностные операторы второго порядка, а I_x, I_y — операторы «правила Симпсона», определенные равенством

$$I_x U_{ij} = \frac{1}{6}[U_{i-1j} + 4U_{ij} + U_{i+1j}]$$

и аналогично для I_y. Приближенное решение в узловых точках, изображенных на рис. 8, приводится в табл. 1; вследствие симметрии необходимо изображать только одну восьмую области. Точное решение в этой задаче есть

$$u(x, y) = \\ = -\left(\frac{\pi}{2}\right)^2 + x^2 + \frac{8}{\pi}\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{(2k-1)^3} \frac{\operatorname{ch}(2k-1)\,y}{\operatorname{ch}((2k-1)\,\pi/2)} \cos(2k-1)\,x.$$

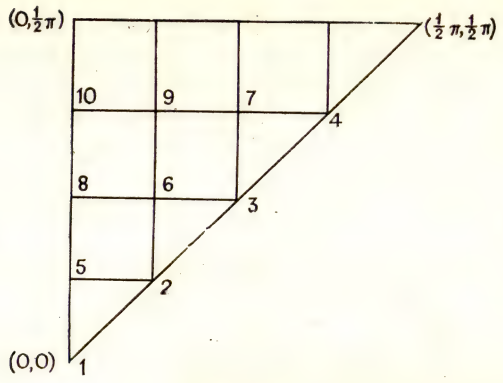


Рис. 8.

Если уравнение (3.2) снова решается в области $-\frac{\pi}{2} < x, y < \frac{\pi}{2}$, но заданы естественные граничные условия

$$\frac{\partial u\left(x, \pm\frac{\pi}{2}\right)}{\partial y} = 0 \quad \left(|x| \leqslant \frac{\pi}{2}\right),$$

$$\frac{\partial u\left(\pm\frac{\pi}{2}, y\right)}{\partial x} = 0 \quad \left(|y| \leqslant \frac{\pi}{2}\right),$$

пространство допустимых функций $\mathcal{H}$ содержит также функции, не равные нулю на границе. Аппроксимирующее подпространство также должно содержать такие функции, поэтому мы берем дополнительные базисные функции $\varphi_{ij}(x, y)$, соот-

Таблица 1

	Решение			
Точка	$T=3$	$T=7$	$T=15$	Точное
1	−1.534	−1.473	−1.459	−1.454
2		−1.321	−1.308	−1.304
3	−0.950	−0.907	−0.897	−0.894
4		−0.370	−0.362	−0.359
5		−1.394	−1.381	−1.376
6		−1.089	−1.078	−1.075
7		−0.566	−0.559	−0.556
8	−1.278	−1.146	−1.135	−1.132
9		−0.666	−0.660	−0.658
10		−0.698	−0.692	−0.690

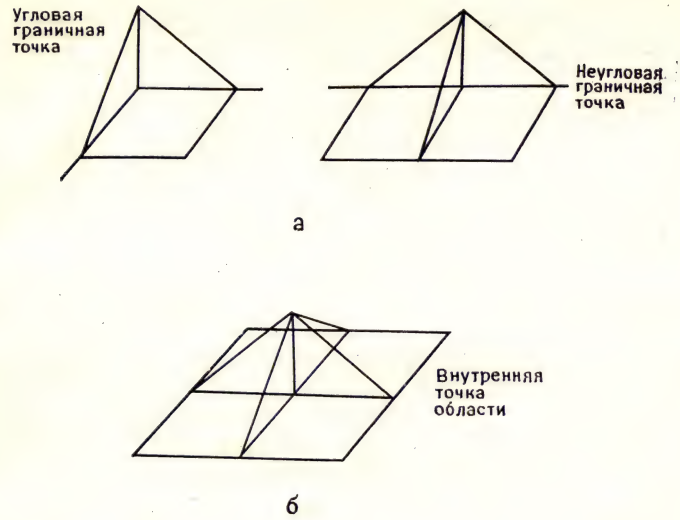


Рис. 9.

ветствующие граничным точкам (x_i, y_j) с $x_i, y_j = \pm \frac{\pi}{2}$. Такие функции не равны нулю самое большее на двух элементах, как показано на рис. 9(a), остальные же базисные функции отличны от нуля на четырех элементах (см. рис. 9(b)).

Упражнение 1. Используя метод Ритца и кусочно-билинейные базисные функции, вычислите коэффициенты уравнения

$$\frac{\partial}{\partial U_{ij}} I = 0,$$

предполагая, что точка (x_i, y_j) есть (a) угловая точка и (*b*) точка на стороне. Здесь берутся естественные граничные условия, $I(v)$ задан в (3.3).

Системы уравнений

Метод Ритца можно применять и к решению систем дифференциальных уравнений с частными производными в теории упругости при малых деформациях. Вообще, если мы берем приближения вида

$$U = \sum_{l=1}^{N_1} \alpha_l \varphi_l, \quad V = \sum_{l=1}^{N_2} \beta_l \psi_l \tag{3.6}$$

и ищем стационарную точку функционала $I(U, V)$ относительно U и V, то получаем $N_1 + N_2$ уравнений

$$\begin{aligned} \frac{\partial}{\partial \alpha_i} I\left(\sum_{j=1}^{N_1} \alpha_j \varphi_j, \sum_{k=1}^{N_2} \beta_k \psi_k\right) = 0 \quad (i = 1, \ldots, N_1), \\ \frac{\partial}{\partial \beta_l} I\left(\sum_{j=1}^{N_1} \alpha_j \varphi_j, \sum_{k=1}^{N_2} \beta_k \psi_k\right) = 0 \quad (l = 1, \ldots, N_2). \end{aligned} \tag{3.7}$$

3.2. Граничные условия

Мы показали, как методом конечных элементов могут быть решены задачи двух различных типов. Вначале мы рассмотрели однородные граничные условия Дирихле $u = 0$. В этом случае все допустимые функции должны удовлетворять этим условиям. Затем мы рассмотрели однородные граничные условия Неймана $\partial u/\partial n = 0$. Здесь на допустимые функции никаких ограничений не накладывается, поскольку граничные условия являются естественными для функционала (3.3).

Граничные условия Дирихле

В общем случае граничные условия Дирихле необходимо вводить как дополнительные условия на аппроксимирующие функции. Это легко сделать если нам нужно решить в некоторой области $R \subset \mathbb{R}^m$ линейное дифференциальное уравнение

$$Au = f \tag{3.8}$$

при условии, что

$$u = g$$

на границе ∂R, причем g является гладкой функцией, допускающей явное аналитическое продолжение внутрь R, т. е. задана такая гладкая функция w, что

$$w = g$$

на ∂R, а Aw определена всюду в R. В этом случае можно рассматривать приближенное решение вида

$$U = w + \sum_{i=1}^{N} \alpha_i \varphi_i, \tag{3.9}$$

где все φ_i равны нулю на ∂R (Синж, 1957). Поскольку функции вида (3.9) образуют линейное многообразие, а не линейное пространство, часто для математического анализа удоб-

нее слегка изменить задачу. После такого изменения задача аппроксимации приобретает структуру классических приближений, описанных в разд. 1.3 и 1.5. Следует подчеркнуть, что численных расчетов это преобразование не затрагивает; меняется лишь математическая формулировка задачи. Следующее упражнение иллюстрирует это.

Упражнение 2. Пусть $f = 0$, $g = -\frac{1}{2}(x^2 + y^2)$, $A = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}$ и уравнение (3.8) рассматривается в области $-\frac{\pi}{2} < x, y < \frac{\pi}{2}$. Пусть

$$w(x, y) = g(x, y) = -\frac{1}{2}(x^2 + y^2). \tag{3.10}$$

Решите уравнение

$$A\left[v(x, y) - \frac{1}{2}(x^2 + y^2)\right] = 0,$$

т. е.

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} - 2 = 0 \tag{3.11}$$

с условием $v = 0$ на границе. Используя решение этого уравнения, заданное табл. 1 или любым другим способом, вычислите решение в узловых точках, изображенных на рис. 8, и сравните это приближение с точным решением

$$u(x, y) = -\left(\frac{\pi}{2}\right)^2 - \frac{1}{2}(y^2 - x^2) + \\ + \frac{8}{\pi}\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{(2k-1)^3} \frac{\operatorname{ch}(2k-1)y}{\operatorname{ch}(2k-1)\pi/2} \cos(2k-1)x.$$

Другой подход к аппроксимации граничных значений

К сожалению, во многих практических задачах граничные условия Дирихле недостаточно гладки, чтобы гарантировать возможность явного аналитического продолжения. В таких задачах может оказаться удобным как-то аппроксимировать граничные данные, используя базисные функции φ_i, не обращающиеся в нуль на границе. Таким образом, мы получаем приближенное решение вида

$$U = \sum_{i=1}^{N} \alpha_i \varphi_i,$$

где некоторые из α_i фиксируются начальными данными. Их можно подобрать, например, так, что приближенное решение

интерполирует граничные условия. Остальные параметры, соответствующие внутренним узлам, вычисляются по методу Ритца. Приближения такого типа не укладываются в рамках классической формулировки метода конечных элементов посредством вычисления вариаций, однако ниже, в гл. 5 показано, что, если все сделано правильно, точность аппроксимации не уменьшается. Широкое признание получил другой подход (Брамбл и Шатц, 1970; Бабушка, 1973; Нитше, 1971), в котором неоднородные граничные условия Дирихле вводятся с помощью так называемого функционала штрафа. Включая граничные условия в функционал, можно удалить все ограничения на аппроксимирующее подпространство K_N. Например, если даны дифференциальное уравнение (3.2) и граничное условие $u = g$, можно использовать функционал

$$J_\lambda(v) = \iint_R \left\{ \frac{1}{2}\left[\left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right] + 2v \right\} dx\,dy + \\ + \frac{1}{2}\lambda \int_{\partial R} (v-g)^2\, d\sigma. \quad (3.12)$$

Ясно, что, если ограничений на аппроксимирующие функции нет, стационарная точка функционала (3.12) соответствует решению уравнения (3.2), удовлетворяющему етественному граничному условию

$$u + (\lambda^{-1})\frac{\partial u}{\partial n} = g$$

на ∂R. Неоднородные граничные условия Дирихле еще встретятся позднее в этой главе, когда будет описываться метод наименьших квадратов.

Неоднородные граничные условия Неймана или смешанные граничные условия могут быть включены в функционал аналогично (3.12) без внесения ограничений на аппроксимирующее подпространство.

3.3. Метод Канторовича (или полудискретный метод)

Другим подходом к применению прямых методов аппроксимации является получение приближенного решения вида

$$U = \sum_{i=1}^{N} \alpha_i \varphi_i,$$

где неизвестные коэффициенты α_i $(i = 1, \ldots, N)$ уже не скаляры, а функции одной из независимых переменных, например, x_1; тогда φ_i рассматриваются как функции оставшихся

m — 1 переменных, т. е.

$$U(x_1, x_2, \ldots, x_m) = \sum_{i=1}^{N} \alpha_i(x_1)\,\varphi_i(x_2, x_3, \ldots, x_m).$$

Этот метод часто связывают с именем Л. В. Канторовича (Канторович, 1933); он лежит в основе большинства конечно-элементных методов решения нестационарных задач, рассмотренных в гл. 6. В применении к краевым задачам метод Канторовича весьма похож на хорошо известный метод прямых (Березин и Жидков, 1962).

Рассмотрим метод Канторовича на простом примере получения приближенного решения уравнения (3.2) с граничными условиями (3.2а). Берется аппроксимирующее подпространство K_N, содержащее функции только одного аргумента y; по (3.2а) такие функции будут удовлетворять условиям

$$\varphi_i\left(-\frac{\pi}{2}\right) = \varphi_i\left(\frac{\pi}{2}\right) = 0 \quad (i = 1, 2, \ldots, N).$$

Приближенное решение вида

$$V(x, y) = \sum_{i=1}^{N} \alpha_i(x)\,\varphi_i(y) \tag{3.13}$$

вычисляется путем минимизации функционала $I(v)$ из (3.3) относительно неопределенных функций α_i $(i = 1, 2, \ldots, N)$. Чтобы сделать это, необходимо сначала переписать функционал I следующим образом:

$$I\left(\sum_{i=1}^{N} \alpha_i\varphi_i\right) = J(\alpha_1, \alpha_2, \ldots, \alpha_N) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} F(\alpha_1, \ldots, \alpha_N)\,dx,$$

где

$$F(\alpha_1, \ldots, \alpha_N) = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{1}{2}\left\{\alpha_i\alpha_j \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{d\varphi_i}{dy}\frac{d\varphi_j}{dy}\,dy + \right.$$

$$\left. + \frac{d\alpha_i}{dx}\frac{d\alpha_j}{dx} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \varphi_i\varphi_j\,dy\right\} + \sum_{i=1}^{N} \alpha_i \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2\varphi_i\,dy. \tag{3.14}$$

Это можно сделать, так как функции φ_i $(i = 1, 2, \ldots, N)$ известны, а интегралы в (3.14) можно вычислить точно. Чтобы получить стационарную точку функционала $J(\alpha_1, \ldots, \alpha_N)$,

мы поступим согласно процедуре, описанной в предыдущей главе, и получим уравнения Эйлера — Лагранжа, соответствующие вариациям каждой $\alpha_i(x)$ $(i = 1, \ldots, N)$. Это необходимо делать с учетом того, что коэффициенты α_i уже не просто скаляры. Поэтому они определяются из системы N уравнений

$$\frac{\partial F}{\partial \alpha_i} - \frac{d}{dx}\left(\frac{dF}{d\alpha_i'}\right) = 0 \quad (i = 1, \ldots, N),$$

где штрих означает дифференцирование по x, а $F(\alpha_1, \ldots, \alpha_N)$ определяется в (3.14). Поэтому неизвестные функции α_i $(i = 1, \ldots, N)$, входящие в приближенное решение (3.13), получаются из системы обыкновенных дифференциальных уравнений

$$\sum_{j=1}^{N} (\alpha_j c_{ij} - \alpha_j'' d_{ij}) = b_i \quad (i = 1, \ldots, N), \tag{3.15}$$

где

$$c_{ij} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{d\varphi_i}{dy}\frac{d\varphi_j}{dy}\,dy,$$

$$d_{ij} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \varphi_i \varphi_j\,dy,$$

$$b_i = -\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2\varphi_i\,dy,$$

при условии, что $\alpha_i\left(-\frac{\pi}{2}\right) = \alpha_i\left(\frac{\pi}{2}\right) = 0$ $(i = 1, \ldots, N)$.

Упражнение 3. Решите систему уравнений (3.15), используя кусочно-линейные базисные функции φ_i $(i = 1, \ldots, N)$, определенные на интервале $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ с равномерным разбиением по формулам (1.3). Сравните результаты, полученные для различных значений N, с приведенными в табл. 1.

Полудискретный метод можно применять к задачам как с неоднородными граничными условиями Дирихле, так и с естественными граничными условиями. Процедура усложняется, когда функционал нужно дополнить интегралами по границе.

Упражнение 4. Найдите функционал $J(\alpha_1, \ldots, \alpha_N)$, который можно было бы использовать для получения решения уравнения (3.2) в области $-\frac{\pi}{2} < x, y < \frac{\pi}{2}$ при условии $u = g$ на границе.

Вообще говоря, при решении краевых задач полудискретный метод эффективен только при условии, что получающаяся одномерная задача может быть решена непосредственно и точно. Несколько таких примеров есть у Эльсгольца (1958), с. 152. Применения к задачам с начальными данными имеют большее значение, и мы будем иметь с ними дело в гл. 6.

3.4. Метод Галеркина

Пока что в этой главе мы рассмотрели ряд методов аппроксимации решения $u(\mathbf{x})$ линейного дифференциального уравнения

$$Au = f \tag{3.16}$$

для $\mathbf{x} \in \mathbb{R}^m$, удовлетворяющего определенным граничным условиям. Для всех этих методов предполагалось, что дифференциальный оператор A удовлетворяет таким условиям, что функция $u(\mathbf{x})$ является решением вариационной задачи

$$\delta I(u) = 0 \tag{3.17}$$

для некоторого функционала $I(u)$. В этом случае можно показать (Михлин, 1976), что функционал может быть записан в виде

$$I(u) = \tfrac{1}{2}(Au, u) - (u, f), \tag{3.18}$$

где

$$(u, v) = \int\limits_R \int u(\mathbf{x})\, v(\mathbf{x})\, d\mathbf{x}.$$

Обобщение этого результата на случай нелинейных операторов подробно рассмотрено Вайнбергом (1956). В нескольких примерах, приведенных в предыдущих разделах этой главы, был использован оператор $A = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}$ вместе с функционалом

$$I(u) = \int\limits_R \int \tfrac{1}{2}(u_x^2 + u_y^2)\, dx\, dy - \int\limits_R \int uf\, dx\, dy. \tag{3.19}$$

Используя теорему Грина (Курант и Гильберт, 1951, с. 241), это выражение можно получить (при определенных граничных условиях) из стандартной формы (3.18).

Упражнение 5. Укажите граничные условия для $u(x, y)$, при которых справедлив переход от (3.18) к (3.19).

Использование интегрирования по частям — т. е. теоремы Грина — для преобразования функционала из стандартной формы (3.18) к такому виду, при котором требуется меньшая гладкость допустимых функций $u(\mathbf{x})$, является одной из основ успеха метода конечных элементов (Прентер, 1975, с. 229).

Аппроксимация Ритца $U(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \varphi_i(\mathbf{x})$ решения вариационной задачи (3.17) может считаться решением уравнения

$$\frac{\partial}{\partial \alpha_i} I(U) = (AU, \varphi_i) - (\varphi_i, f) = 0 \quad (i = 1, \ldots, N) \tag{3.20}$$

только тогда, когда мы имеем оператор A требуемого вида. Но даже если это не так и (3.20) не справедливо, система уравнений

$$(AU, \varphi_i) - (\varphi_i, f) = (AU - f, \varphi_i) = 0 \quad (i = 1, \ldots, N) \tag{3.21}$$

все еще определяет приближенное решение $U(\mathbf{x})$ для (3.16). В действительности систему (3.21) можно использовать, даже если оператор нелинеен. Таким образом, метод конечных элементов можно использовать для решения значительно более широкого класса задач, чем класс задач, допускающих вариационную формулировку. Однако все же желательно, чтобы (3.21) можно было бы интегрированием по частям преобразовать так, чтобы уменьшить требуемую гладкость функций φ_i. Такие аппроксимации решения обычно связывают с именем русского математика Б. Г. Галеркина (1871—1945). Во многих советских учебниках о них говорят как о методе Бубнова — Галеркина (Михлин и Смолицкий, 1965; Вулих, 1967).

Слабые решения

Часто функцию $u(\mathbf{x})$, которая удовлетворяет линейному дифференциальному уравнению

$$Au = f \quad (\mathbf{x} \in R),$$

называют классическим решением, в отличие от слабого решения, которое удовлетворяет уравнению

$$(Au, v) = (f, v) \quad \text{(для всех } v \in \mathcal{H}) \tag{3.22a}$$

или

$$(u, A^* v) = (f, v) \quad \text{(для всех } v \in \mathcal{H}_1). \tag{3.22b}$$

Пространство $\mathcal{H}$ содержит все измеримые допустимые функции, которые обращаются в нуль на границе ∂R; про такие функции иногда говорят, что они имеют *компактный носитель*. Пространство $\mathcal{H}_1 \subset \mathcal{H}$ содержит только те допустимые функции, для которых $A^* v$ измерима.

Отсюда следует, что если оператор A имеет порядок $2k$, то слабая форма (3.22a) требует, чтобы решение u имело измеримые производные порядка $2k$, что в терминах пространства Соболева (см. гл. 5) можно записать как $u \in \mathscr{H}_2^{(2k)}(R)$. Однако вторая слабая форма задачи (3.22b) требует только, чтобы $u \in \mathscr{L}_2(R)$; для *пробных функций* v — наоборот. Дополнительное требование обращения в нуль v на границе обычно — как в гл. 5 — указывается как $v \in \mathring{\mathscr{H}}_2^{(2k)}(R)$.

С вычислительной точки зрения наиболее полезная слабая форма задачи, называемая далее *галеркинской формой* [1]), возникает из (3.22a) или из (3.22b) после k интегрирований по частям. В этой галеркинской форме требования гладкости минимальны в том смысле, что $u \in \mathscr{H}_2^{(k)}(R)$, $v \in \mathring{\mathscr{H}}_2^{(k)}(R)$. Обычно в этом случае задача записывается так:

$$a(u, v) = (f, v) \quad \text{(для всех } v \in \mathscr{H}_2\text{)}, \tag{3.23}$$

где новое пространство пробных функций $\mathscr{H}_2$, очевидно, удовлетворяет вложениям

$$\mathscr{H}_1 \subset \mathscr{H}_2 \subset \mathscr{H}.$$

Например, галеркинская форма дифференциального оператора $-\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}$ имеет вид

$$a(u, v) = \iint_R \left[\left(\frac{\partial u}{\partial x} \right) \left(\frac{\partial v}{\partial x} \right) + \left(\frac{\partial u}{\partial y} \right) \left(\frac{\partial v}{\partial y} \right) \right] dx\, dy,$$

где в этом случае $k = 1$. Форма Галеркина дифференциального оператора $\frac{\partial^4}{\partial x^4} + 2\frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4}$ (см. гл. 2) выглядит так:

$$a(u, v) = \iint_R \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) \left(\frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial y^2} \right) dx\, dy.$$

Можно показать, что при достаточно внимательном отношении к граничным условиям слабое решение единственно и совпадает с классическим: один из таких результатов приводится в теореме 5.2 гл. 5. Существование и единственность слабых решений изучаются также в книгах Берса, Джона и Шехтера (1966), с. 206, Нечаса (1967), с. 23 и Лионса и Мадженеса (1971), § 2.9. В этой книге предполагается, что пространства $\mathscr{H}_1$, $\mathscr{H}_2$, $\mathscr{H}$ выбраны правильно, а тогда решения задач (3.22a), (3.22b) и (3.23) совпадают с требуемым классическим решением. Вообще единственной рассматриваемой

[1]) Под формой Галеркина здесь понимаются и слабая постановка задачи, и слабая форма уравнения и некоторый билинейный функционал. Нам кажется, что в контексте это не вызовет недоразумения.—*Прим. перев.*

здесь слабой формой является форма Галеркина. Аппроксимация Галеркина U удовлетворяет системе

$$a(U, \varphi_i) = (f, \varphi_i) \quad (i = 1, \ldots, N). \tag{3.24}$$

Поскольку функции φ_i порождают конечномерное подпространство K_N, утверждение, эквивалентное системе уравнений (3.24), есть

$$a(U, V) = (f, V) \quad (\text{для всех } V \in K_N). \tag{3.24а}$$

Заметим, что u (или U) не обязательно из энергетического пространства, это требование накладывается только на v (или V). Так, если заданы неоднородные граничные условия Дирихле, то может быть удобным использовать аппроксимацию вида

$$U(\mathbf{x}) = W(\mathbf{x}) + \sum_{i=1}^{N} \alpha_i \varphi_i(\mathbf{x}),$$

где функция W удовлетворяет граничным условиям.

В уравнениях Галеркина естественно предполагать, что приближение U и пробные функции V определяются одним множеством базисных функций φ_i $(i = 1, \ldots, N)$. В этом случае в задачах, в которых возможны как аппроксимации по Ритцу, так и аппроксимации по Галеркину, оба метода приводят к одному и тому же решению. Другая аппроксимация получается с использованием тех же базисных функций $\varphi_i \in K_N$, но при пробных функциях, выраженных через некоторые $\psi_i \in L_N$ $(i = 1, \ldots, N)$, где K_N и L_N — разные подпространства пространства $\mathcal{H}$.

Метод наименьших квадратов можно рассматривать как метод этого типа с $\psi_i = A\varphi_i$ $(i = 1, \ldots, N)$. Использование двух множеств, $\{\varphi_i\}$ и $\{\psi_i\}$, несовпадающих базисных функций для определения аппроксимации указанного выше вида из условий

$$a(U, \psi_i) = (f, \psi_i) \quad (i = 1, \ldots, N) \tag{3.25}$$

приводит различных авторов к развитию так называемых *сопряженных аппроксимаций* (см. Оден, 1976, и приведенную там библиографию) и к так называемому *методу взвешенных невязок* (см. Финлейсон и Скривен, 1967, и имеющиеся там ссылки). В разд. 7.4(Е) на примере показано, что в некоторых практических задачах есть определенные вычислительные преимущества, связанные с выбором пробных функций другого вида, нежели приближенное решение. Однако ни здесь, ни в гл. 5 нет возможности для обсуждения теоретических выгод таких методов; для этого читателю следует обратиться к указанной литературе.

Метод Галеркина можно применить к решению нелинейных задач, но только в специальных случаях удается получить слабую форму с меньшими требованиями гладкости. Один из примеров такого рода — дифференциальное уравнение

$$\frac{\partial}{\partial x}\left(p(u)\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(q(u)\frac{\partial u}{\partial y}\right)+f(x, y)=0.$$

Оно приводит к форме Галеркина

$$a(U, \varphi_i)=\iint_R\left[p(U)\left(\frac{\partial U}{\partial x}\right)\left(\frac{\partial \varphi_i}{\partial x}\right)+q(U)\left(\frac{\partial U}{\partial y}\right)\left(\frac{\partial \varphi_i}{\partial y}\right)\right]dx\,dy= \\ =(f, \varphi_i) \quad (i=1, 2, \ldots, N)$$

Сопряженные формулировки

Вариационное исчисление можно распространить на широкий класс задач, допускающих применение метода Галеркина. В предыдущей главе был получен формальный лагранжиан *диссипативной системы* путем рассмотрения сопряженной задачи. Если ищутся приближенные решения как *основной*, так и *сопряженной задач*, то аппроксимация Галеркина находится из условия стационарности этого лагранжиана. Фактически (3.24) возникает как необходимое условие стационарности лагранжиана.

В качестве простого примера рассмотрим уравнение

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+2\frac{\partial u}{\partial y}=0 \tag{3.26}$$

в некоторой области $R \subset \mathbb{R}^2$ при условии $u=0$ на границе ∂R. Будем называть эту задачу основной. Сопряженной задачей будет решение сопряженного уравнения

$$\frac{\partial^2 u^*}{\partial x^2}+\frac{\partial^2 u^*}{\partial y^2}-2\frac{\partial u^*}{\partial y}=0 \tag{3.27}$$

в R при аналогичном граничном условии $u^*=0$.

Нетрудно убедиться в том, что (3.26) и (3.27) оказываются уравнениями Эйлера — Лагранжа, соответствующими вариации функционала

$$I(u, u^*)=\iint_R\left[\frac{\partial u}{\partial x}\frac{\partial u^*}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial u^*}{\partial y}+u\frac{\partial u^*}{\partial y}-u^*\frac{\partial u}{\partial y}\right]dx\,dy. \tag{3.28}$$

Следует, однако, помнить, что основное уравнение (3.26) получается из условия

$$\delta_{u^*} I(u, u^*)=0, \tag{3.29а}$$

а не из условия

$$\delta_u I(u, u^*), \tag{3.29b}$$

так как (3.29b) ведет к сопряженному уравнению.

Если мы ищем приближенное решение основной задачи в виде $U = \sum_{i=1}^{N} \alpha_i \varphi_i$, то приближенное решение сопряженной задачи необходимо получить в той же форме, $U^* = \sum_{i=1}^{N} \beta_i \varphi_i$; другой вид сопряженной аппроксимации даст скорее (3.25), чем (3.24). Эти приближения можно найти как стационарную точку функционала (3.28) относительно неизвестных коэффициентов α_i и β_i. Таким образом мы от (3.29a) приходим к

$$\frac{\partial}{\partial \beta_i} I(U, U^*) = 0 \quad (i = 1, \ldots, N), \tag{3.30a}$$

а от (3.29b) — к

$$\frac{\partial}{\partial \alpha_i} I(U, U^*) = 0 \quad (i = 1, \ldots, N). \tag{3.30b}$$

Если мы рассматриваем функционал $I(U, U^*)$ из (3.28), приближенное решение основного уравнения получается из системы уравнений

$$\iint_R \left\{ \left(\frac{\partial U}{\partial x}\right)\left(\frac{\partial \varphi_i}{\partial x}\right) + \left(\frac{\partial U}{\partial y}\right)\left(\frac{\partial \varphi_i}{\partial y}\right) + U\left(\frac{\partial \varphi_i}{\partial y}\right) - \left(\frac{\partial U}{\partial y}\right)\varphi_i \right\} dx\, dy = 0 \tag{3.31}$$

$$(i = 1, \ldots, N).$$

Более прямой способ получения аппроксимации Галеркина следует из слабой формы уравнения (3.26). Легко проверить, что слабой формой Галеркина уравнения (3.26) является

$$a(u, v) = 0 \quad (\text{для всех } v \in \mathcal{H}),$$

где

$$a(u, v) = \iint_R \left\{ \left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial v}{\partial x}\right) + \left(\frac{\partial u}{\partial y}\right)\left(\frac{\partial v}{\partial y}\right) + u\left(\frac{\partial v}{\partial y}\right) - \left(\frac{\partial u}{\partial y}\right) v \right\} dx\, dy,$$

а отсюда сразу получаем, что аппроксимация Галеркина, возникающая из системы уравнений

$$a(U, \varphi_i) = 0 \quad (i = 1, \ldots, N), \tag{3.32}$$

удовлетворяет уравнениям (3.31). Таким образом, на примере данной задачи показано, как одна и та же система уравнений может быть получена двумя математически разными спо-

собами: как слабая форма исходного дифференциального уравнения, или как условие стационарности некоторого функционала. Сопряженная задача, нужная для определения подходящего функционала, вообще говоря, является всего лишь математическим приемом и не имеет физического смысла. Поэтому в дальнейшем метод Галеркина исследуется на основе слабой формы уравнения, на другую же формулировку делаются только ссылки.

В качестве примера возьмем уравнение (3.26) в области $-\frac{1}{2}\pi < x,\ y < \frac{1}{2}\pi$ с граничными условиями

$$\begin{aligned} u\left(\pm\tfrac{1}{2}\pi, y\right) &= 0 \qquad \left(|y| \leqslant \tfrac{1}{2}\pi\right), \\ u\left(x, -\tfrac{1}{2}\pi\right) &= 0 \qquad \left(|x| \leqslant \tfrac{1}{2}\pi\right), \\ u\left(x, \tfrac{1}{2}\pi\right) &= \left(\tfrac{\pi}{2}\right)^2 - x^2 \quad \left(|x| \leqslant \tfrac{1}{2}\pi\right). \end{aligned} \tag{3.33}$$

Применим к этой задаче метод конечных элементов Галеркина для получения приближенного решения на классе кусочно-билинейных функций, использованном ранее. Введем функцию

$$w(x, y) = \left[\left(\tfrac{\pi}{2}\right)^2 - x^2\right]\left(\tfrac{\pi}{2} + y\right)\tfrac{1}{\pi},$$

которая удовлетворяет граничным условиям, после чего найдем аппроксимацию Галеркина вида

$$U(x, y) = \sum_{i,\,j=1}^{T} \alpha_{ij}\varphi_{ij}(x, y) + w(x, y)$$

из системы уравнений (3.31). Результаты вычислений при различных сетках, описанных ранее, приводятся в табл. 2. На рис. 10 изображено положение точек, включенных в таблицу. Хотя здесь нет той симметрии, которая была ранее, необхо-

Таблица 2

	Решение			
Точка	$T = 3$	$T = 7$	$T = 15$	Точное
1	1.826	1.822	1.821	1.821
2	1.324	1.314	1.311	1.310
3	0.934	0.870	0.859	0.855
4	1.301	1.309	1.310	1.311
5	0.940	0.933	0.931	0.931
6	0.665	0.617	0.608	0.605

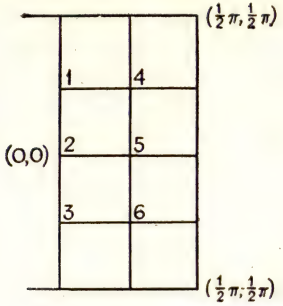


Рис. 10.

димо рассматривать все еще только половину области. Точное решение уравнения (3.26), удовлетворяющее граничным условиям (3.33), выглядит так:

$$u(x, y) = \frac{4}{\pi} e^{\frac{\pi}{2} - y} \sum_{k=1}^{\infty} \left\{ \frac{\sin([2k-1]^2 + 1)^{1/2} y}{\sin([2k-1]^2 + 1)^{1/2} \frac{\pi}{2}} + \right.$$
$$\left. + \frac{\operatorname{ch}([2k-1]^2 + 1)^{1/2} y}{\operatorname{ch}([2k-1]^2 + 1)^{1/2} \frac{\pi}{2}} \right\} \frac{\cos(2k-1)x}{(-1)^{k-1}(2k-1)^3}.$$

Здесь следует подчеркнуть, что во всех вариантах метода Галеркина система уравнений получается без особых затруднений.

Полудискретный метод Галеркина

Если метод Галеркина следует из вариационного принципа, получаемого с помощью введения сопряженной задачи, должно быть ясно, что, применяя метод, описанный в разд. 3.3, можно свести диссипативное уравнение с частными производными к системе обыкновенных дифференциальных уравнений. Такой подход имеет небольшую ценность для решения краевых задач, и так как для задач с начальными данными нет вариационной формулировки, мы не рассматриваем подробно этот метод. Следующее ниже упражнение предназначено читателям, интересующимся данным методом.

Упражнение 6. Положим

$$w(x, y) = \sum_{i=1}^{2} \alpha_i(x) \varphi_i(y); \tag{3.34a}$$

$$w^*(x, y) = \sum_{i=1}^{2} \beta_i(x) \varphi_i(y), \tag{3.34b}$$

где φ_1 и φ_2 — кусочно-линейные функции. Постройте функционал $I(w, w^*)$, соответствующий уравнению (3.26) с граничными условиями (3.33), а затем, используя $w(x, y)$ и $w^*(x, y)$, указанные в (3.34a) и (3.34b), найдите решение уравнений

$$\delta_{\beta_i} J(\alpha_1, \alpha_2, \beta_1, \beta_2) = 0 \quad (i = 1, 2),$$

т. е. определите полудискретное решение Галеркина уравнения (3.26), удовлетворяющее (3.33). Сравните ваше решение с результатами, приведенными в табл. 2.

Метод переменных направлений Галеркина (ПНГ) для случая прямоугольных областей (Дуглас и Дюпон, 1971)

Когда метод конечных элементов применяется к одномерным линейным задачам, матрица получающейся системы линейных алгебраических уравнений имеет простой *ленточный* вид, тогда как задачи большей размерности дают *блочно-ленточные* матрицы, у которых каждый блок сам является ленточной матрицей. Например, в двумерном случае билинейная аппроксимация часто приводит к системе с матрицей вида

$$G = \begin{bmatrix} DE & & & \\ CDE & & & \\ \cdot\cdot\cdot & & & \\ & \cdot\cdot\cdot & & \\ & & \cdot\cdot\cdot & \\ & & CDE & \\ & & CD & \end{bmatrix},$$

где C, D и E — трехдиагональные матрицы.

Имеется несколько алгоритмов решения систем уравнений с ленточными матрицами, эффективных и дешевых (т. е. дающих максимальную точность при минимуме времени и памяти), однако такие методы менее эффективны и значительно дороже в применении к задачам с блочно-ленточными матрицами. Цель метода ПНГ в методе конечных элементов та же, что и в методе переменных направлений в разностных схемах, а именно свести систему уравнений многомерной задачи к последовательности систем, по форме аналогичных системам уравнений, возникающих в одномерных задачах (Митчелл, 1967).

Для того чтобы применять метод ПНГ, необходимо предполагать, что базисные функции имеют *форму тензорного произведения*, т. е.

$$\varphi_{ij}(x, y) = \varphi_i(x)\,\varphi_j(y) \quad (i = 1, \ldots, N_x;\ j = 1, \ldots, N_y).$$

Матрицу $G = \{a(\varphi_{ij}, \varphi_{kl})\}$ можно тогда разложить так, что

$$a(\varphi_{ij}, \varphi_{kl}) = a_x(\varphi_i, \varphi_k)\, b_y(\varphi_j, \varphi_l) + b_x(\varphi_i, \varphi_k)\, a_y(\varphi_j, \varphi_l)$$
$$(i, k = 1, \ldots, N_x;\ j, l = 1, \ldots, N_y).$$

Если A — $(N \times N)$-матрица и B — $(M \times M)$-матрица, то матричное тензорное произведение, обозначаемое $A \otimes B$, есть $(NM \times NM)$-матрица

$$\begin{bmatrix} a_{11}B \ldots a_{1N}B \\ \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \\ a_{N1}B \ldots a_{NN}B \end{bmatrix}.$$

Теперь матрицу G можно записать как

$$G = A_x \otimes B_y + B_x \otimes A_y,$$

если узлы упорядочены по столбцам. Система уравнений принимает вид

$$\{A_x \otimes B_y + B_x \otimes A_y\}\, \alpha = \mathbf{b}$$

и решается посредством итераций

$$(\lambda_n B_x + A_x) \otimes (\lambda_n B_y + A_y)\, \alpha^{(n)} =$$
$$= (\lambda_n B_x - A_x) \otimes (\lambda_n B_y - A_y)\, \alpha^{(n-1)} + 2\lambda_n \mathbf{b} = \psi^{(n-1)}$$

в два этапа:

$$(\lambda_n B_x + A_x) \otimes I_{N_y} \alpha^{(n^*)} = \psi^{(n-1)}, \tag{3.35a}$$

$$I_{N_x} \otimes (\lambda_n B_y + A_y)\, \alpha^{(n)} = \alpha^{(n^*)}. \tag{3.35b}$$

Можно расщепить эти уравнения так, что если $\alpha_{p,\,C} = (\alpha_{p1}, \ldots, \alpha_{pN_y})^T$ $(p = 1, \ldots, N_x)$ — столбец значений на сетке, а $\alpha_{p,\,R} = (\alpha_{1p}, \ldots, \alpha_{N_x p})^T$ $(p = 1, \ldots, N_y)$ — строка значений на сетке, то (3.35a) переходит в

$$(\lambda_n B_x + A_x)\, \alpha^{(n^*)}_{p,\,R} = \psi^{(n-1)}_{p,\,R} \quad (p = 1, \ldots, N_y),$$

а (3.35b) — в

$$(\lambda_n B_y + A_y)\, \alpha^{(n)}_{p,\,C} = \alpha^{(n^*)}_{p,\,C} \quad (p = 1, \ldots, N_x).$$

Дуглас и Дюпон показали, что при подходящем выборе последовательности параметров итераций $\{\lambda_n\}$ метод ПНГ оказывается быстро сходящимся.

Метод коллокации

Метод коллокации во многих отношениях подобен методу Галеркина. Он предусматривает такой выбор коэффициентов α_i $(i = 1, \ldots, N)$ в представлении

$$U = \sum_{i=1}^{N} \alpha_i \varphi_i,$$

что дифференциальное уравнение удовлетворяется точно в некоторых определенных точках. Было показано (де Бур и Шварц, 1973; Лукас и Редин, 1972; Элберг и Ито, 1975), что для обыкновенных дифференциальных уравнений при правильном выборе точек коллокации метод подобен по точности методу Галеркина с тем же самым набором базисных функций φ_i $(i = 1, \ldots, N)$. Если, например, базисные функции являются эрмитовыми кусочно-полиномиальными функциями степени $2r - 1$, точки коллокации на кажом подынтервале $[x_i, x_{i+1}]$ берутся в нулях полинома Лежандра $P_r\left(\frac{2x - x_{i+1} - x_i}{x_{i+1} - x_i}\right)$.

Преимущества метода коллокации состоят в следующем:

(I) Нет скалярных произведений, а значит, не нужно интегрировать, как в методах Галеркина и Ритца.

(II) Окончательные алгебраические уравнения имеют меньше членов, чем соответствующие уравнения для аппроксимаций Галеркина.

Главный недостаток метода коллокации заключается в необходимости использовать базисные функции степени (по меньшей мере) $2k$ для дифференциального уравнения порядка $2\,k$.

Методы, использующие коллокацию, были разработаны также и для эволюционных задач (Дуглас и Дюпон, 1973). Пример, включающий коллокацию, приводится в разд. 7.4(В).

3.5. Проекционные методы

Ранее в этой главе эквивалентность вариационного уравнения

$$\delta_v I(v) = 0 \tag{3.36}$$

и дифференциального уравнения

$$Au = f \tag{3.37}$$

использовалась для получения приближенного решения Ритца вида

$$U = \sum_{i=1}^{N} \alpha_i \varphi_i.$$

Параметры α_i $(i = 1, \ldots, N)$ определяются из вариационных уравнений

$$\delta_{\alpha_i} I\left(\sum_{j=1}^{N} \alpha_j \varphi_j\right) = 0 \quad (i = 1, \ldots, N),$$

которые можно записать в более явной форме

$$\sum_{j=1}^{N} a(\varphi_i, \varphi_j)\alpha_j = (\varphi_i, f) \quad (i = 1, \ldots, N). \tag{3.38}$$

В разд. 1.3 показано, что если $\mathcal{H}$ — гильбертово пространство со скалярным произведением $[,]$ и K_N — любое N-мерное подпространство $\mathcal{H}$, то для любого $q \in \mathcal{H}$ его ортогональная проекция на K_N,

$$\tilde{q} = \sum_{i=1}^{N} \alpha_i \varphi_i, \tag{3.39}$$

такова, что

$$\sum_{j=1}^{N} [\varphi_i, \varphi_j]\alpha_j = [\varphi_i, q] \quad (i = 1, \ldots, N), \tag{3.40}$$

где φ_i $(i = 1, \ldots, N)$ образуют базис K_N. Отсюда следует, что $\tilde{q}$ является единственной наилучшей аппроксимацией, как и в разд. 1.2. Теперь покажем, что не только (3.38) и (3.40) кажутся похожими, но и аппроксимации обладают аналогичными свойствами.

Пусть $\mathcal{H}$ — пространство допустимых функций, соответствующее дифференциальному уравнению $Au = f$ с заданными граничными условиями. Тогда *энергетическое пространство* $\mathcal{H}_A$ оператора A данной задачи есть подпространство $\mathcal{H}$, содержащее все допустимые функции $u(\mathbf{x})$ и $v(\mathbf{x})$, такие, что функционал $a(u, v)$, определенный в разд. 3.4, ограничен. Скалярное произведение в энергетическом пространстве $\mathcal{H}_A$ определяется как

$$(u, v)_A = a(u, v),$$

а норма, которую называют *энергетической нормой*, — как

$$\|u\|_A^2 = (u, u)_A.$$

В этом определении предполагается, что если можно преобразовать скалярное произведение $(,)_A$, используя интегри-

рование по частям, то именно преобразованная форма используется в определении энергетического пространства. Если это сделано, энергетическое пространство оказывается полным, а значит, гильбертовым.

Теорема 3.1. *Если линейный оператор A положителен и самосопряжен, приближенное решение Ритца дифференциального уравнения $Au = f$ является ортогональной проекцией точного решения на аппроксимирующее подпространство энергетического пространства. Таким образом, аппроксимация Ритца есть наилучшая аппроксимация в смысле энергетического пространства.*

Доказательство. Доказательство прямо следует из определения аппроксимации Ритца, данного в разд. 3.1. Наилучшая аппроксимация

$$U = \sum_{i=1}^{N} \alpha_i \varphi_i$$

точного решения u_0 в смысле энергетической нормы есть $U \in K_N$, такое, что

$$(U - u_0, \varphi_i)_A = 0 \quad (i = 1, \ldots, N),$$

т. е.

$$a(U - u_0, \varphi_i) = 0 \quad (i = 1, \ldots, N). \tag{3.41}$$

Если решение единственно, то (3.41) можно переписать как (3.38). Отсюда следует требуемый результат, если только a — действительно норма. Можно показать (Михлин, 1976, с. 86), что билинейная форма a является нормой тогда и только тогда, когда (3.36) эквивалентно (3.37), а это завершает доказательство.

Приближенное решение Ритца уравнения (3.2), например, есть наилучшая аппроксимация точного решения в смысле полунормы Дирихле

$$|u|_{1,R} = \left\{ \iint_R \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dx\, dy \right\}^{1/2}.$$

Упражнение 7. (I) Покажите, что для того, чтобы $(,)_A$ было скалярным произведением, а $\|.\|_A$ — нормой, оператор A должен быть положительно определенным и самосопряженным.

(II) Покажите, что $a(u, v)$ ограничена тогда и только тогда, когда и $\|u\|_A$ и $\|v\|_A$ ограничены. Отсюда покажите, что u — элемент энергетического пространства тогда и только тогда, когда $\|u\|_A$ ограничена.

Упражнение 8. Покажите, что функционал $I(v)$, соответствующий дифференциальному уравнению $Au = f$ в области R с граничным условием $u = 0$, можно переписать как

$$I(v) = \| v - u_0 \|_A^2 - \| u_0 \|_A^2 \tag{3.42}$$

или

$$I(v) = \| v \|_A^2 - 2(v, u_0)_A,$$

где u_0 — точное решение дифференциального уравнения.

Упражнение 9. Укажите класс дифференциальных уравнений, для которых аппроксимация Ритца оказывается наилучшей в смысле нормы Соболева:

$$\| u \|_{1, R} = \left\{ \iint\limits_R \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 + u^2 \right] dx\, dy \right\}^{1/2}.$$

Метод наименьших квадратов

Метод Ритца дает наилучшую аппроксимацию решения линейного дифференциального уравнения $Au = f$ в смысле энергетической нормы тогда и только тогда, когда оператор A — положительно определенный и самосопряженный. Хотя метод Галеркина можно использовать для приближенного решения более широкого класса задач, мы уже не получим наилучшей аппроксимации того же типа. Чтобы получить «наилучшую аппроксимацию» в несамосопряженных задачах, необходимо переформулировать процедуру аппроксимации и ввести так называемый метод наименьших квадратов.

Напомним, что аппроксимация Ритца есть решение вариационной задачи

$$\min_{U \in K_N} \| U - u_0 \|_A^2,$$

где u_0 — неизвестное решение. Теоретически можно заменить энергетическую норму $\|.\|_A$ на любую другую при условии, что нам требуется Au_0, а не решение u_0 само по себе. Таким образом, мы можем определить аппроксимацию U как решение задачи

$$\min_{U \in K_N} \| AU - Au_0 \|^2$$

или

$$\min_{U \in K_N} \| AU - f \|^2;$$

это и есть основа метода наименьших квадратов (Брамбл и Шатц, 1970; 1971). Если применить вариационное исчисление к вычислению необходимых условий стационарности такого

функционала, получается уравнение Эйлера — Лагранжа, включающее оператор A^*A, т. е. получается уравнение более высокого порядка, чем исходное. По этой причине необходимо быть осторожными с утверждением, что решения задач совпадают, особенно когда заданы неоднородные условия Дирихле; таким образом, если только не налагаются дополнительные условия на аппроксимирующие функции, мы должны видоизменить функционал (см. разд. 3.2). Если мы полагаем $u = g$ на ∂R, аппроксимация по методу наименьших квадратов могла бы минимизировать, например,

$$\iint_R (AU - f)^2 \, d\mathbf{x} + \lambda \int_{\partial R} (U - g)^2 \, d\sigma$$

при подходящем выборе λ.

Хотя подробное рассмотрение метода наименьших квадратов остается вне рамок этой книги, он снова упоминается в разд. 5.4(D). Численные эксперименты с методом наименьших квадратов можно найти в гл. 7.

ГЛАВА 4

БАЗИСНЫЕ ФУНКЦИИ

В разд. 1.1. были построены элементарные базисные функции для прямоугольных и многоугольных областей, причем первые разбивались на прямоугольные элементы, а вторые — на треугольные. В данной главе базисные функции строятся для разнообразных по форме элементов в случаях двух и трех измерений.

4.1. Треугольник

(А) Лагранжева интерполяция

Треугольник, или двумерный симплекс, является, вероятно, наиболее широко используемым конечным элементом. Одна из причин этого в том, что любую область в двумерном пространстве можно аппроксимировать многоугольниками, которые всегда можно разбить на конечное число треугольников. Кроме того, полный полином порядка m

$$\Pi_m(x, y) = \sum_{k+l=0}^{m} \alpha_{kl} x^k y^l \tag{4.1}$$

может быть использован для интерполяции функции, скажем $U(x, y)$, по значениям в $\frac{1}{2}(m+1)(m+2)$ симметрично расположенных узлах в треугольнике. Первые три случая этого общего представления для треугольника $P_1P_2P_3$ с координатами вершин соответственно (x_1, y_1), (x_2, y_2), (x_3, y_3) таковы:

(1) *Линейный случай* $(m = 1)$. Здесь интерполирующий полином имеет вид

$$\Pi_1(x, y) = \alpha_1 + \alpha_2 x + \alpha_3 y = \sum_{j=1}^{3} U_j p_j^{(1)}(x, y),$$

где U_j $(j = 1, 2, 3)$ — значения $U(x, y)$ в вершинах p_j, а

$$p_j^{(1)}(x, y) = \frac{1}{C_{jkl}}(\tau_{kl} + \eta_{kl} x - \xi_{kl} y) = \frac{D_{kl}}{C_{jkl}}, \tag{4.2}$$

где

$$\tau_{kl} = x_k y_l - y_k x_l, \quad \xi_{kl} = x_k - x_l, \quad \eta_{kl} = y_k - y_l,$$

$$D_{kl} = \det \begin{bmatrix} 1 & x & y \\ 1 & x_k & y_k \\ 1 & x_l & y_l \end{bmatrix}, \quad C_{jkl} = \det \begin{bmatrix} 1 & x_j & y_j \\ 1 & x_k & y_k \\ 1 & x_l & y_l \end{bmatrix},$$

причем (j, k, l) — произвольная перестановка из (1, 2, 3), а $|C_{jkl}|$ — удвоенная площадь треугольника $P_1P_2P_3$; нетрудно заметить, что

$$p_j^{(1)}(x_k, y_k) = \begin{cases} 1 & (j = k) \\ 0 & (j \neq k) \end{cases} \quad (1 \leqslant j, k \leqslant 3).$$

(2) *Квадратичный случай* ($m = 2$). Теперь полином имеет вид

$$\Pi_2(x, y) = \sum_{j=1}^{6} U_j p_j^{(2)}(x, y), \tag{4.3}$$

где U_j $(j = 1, \ldots, 6)$ — значения $U(x, y)$ в вершинах P_j $(j = 1, 2, 3)$ и в средних точках P_j $(j = 4, 5, 6)$ сторон P_1P_2, P_2P_3, P_3P_1 соответственно. Функции $p_j^{(2)}(x, y)$ $(j = 1, 2, \ldots, 6)$ задаются формулами

$$p_1^{(2)}(x, y) = p_1^{(1)}\left(2p_1^{(1)} - 1\right)$$

($p_2^{(2)}(x, y)$ и $p_3^{(2)}(x, y)$ — аналогично) и

$$p_4^{(2)}(x, y) = 4p_1^{(1)}p_2^{(1)}$$

($p_5^{(2)}(x, y)$ и $p_6^{(2)}(x, y)$ — аналогично). Снова видно, что

$$p_j^{(2)}(x_k, y_k) = \begin{cases} 1 & (j = k) \\ 0 & (j \neq k) \end{cases} \quad (1 \leqslant j, k \leqslant 6).$$

Особо отметим, что базисные функции $p_j^{(2)}(x, y)$ $(j = 1, \ldots, 6)$ можно выразить через базисные функции $p_j^{(1)}(x, y)$ $(j = 1, 2, 3)$. Это верно почти для всех базисных функций, которые мы будем рассматривать, и поэтому для упрощения формул мы будем писать просто p_j вместо $p_j^{(1)}$ $(j = 1, 2, 3)$.

(3) *Кубический случай* ($m = 3$). Интерполирующий полином есть

$$\Pi_3(x, y) = \sum_{j=1}^{10} U_j p_j^{(3)}(x, y), \tag{4.4}$$

где U_j $(j = 1, 2, 3)$ — значения $U(x, y)$ в вершинах P_1, P_2, P_3, U_j $(j = 4, 5, \ldots, 9)$ — в точках трисекции сторон, а U_{10} — значение $U(x, y)$ в центре тяжести треугольника. Базисные функ-

ции таковы:

$$p_1^{(3)}(x, y) = \frac{1}{2} p_1 (3p_1 - 1)(3p_1 - 2),$$

аналогично $p_2^{(3)}(x, y)$ и $p_3^{(3)}(x, y)$,

$$p_4^{(3)}(x, y) = \frac{9}{2} p_1 p_2 (3p_1 - 1), \quad p_5^{(3)}(x, y) = \frac{9}{2} p_1 p_2 (3p_2 - 1),$$

аналогично $p_6^{(3)}(x, y), \ldots, p_9^{(3)}(x, y)$, а

$$p_{10}^{(3)}(x, y) = 27 p_1 p_2 p_3.$$

Десятый параметр может быть исключен с помощью линейного соотношения

$$U_{10} = \frac{1}{4} \sum_{j=4}^{9} U_j - \frac{1}{6} \sum_{j=1}^{3} U_j$$

и получающаяся функция будет еще точно интерполировать квадратичные функции (Сьярле и Равьяр, 1972 а); это пример так называемого *исключения внутренних параметров*. Треугольники для $m = 1, 2, 3$ показаны на рис. 11.

Упражнение 1. Проверьте, что

$$U(x, y) = \sum_{j=1}^{9} U_j \bar{p}_j^{(3)}(x, y)$$

интерполирует квадратичные полиномы точно, если $\bar{p}_j^{(3)} = p_j^{(3)} + \alpha_j p_{10}^3$ $(j = 1, \ldots, 9)$, где

$$\alpha_j = \begin{cases} -\frac{1}{6} & (j = 1, 2, 3) \\ \frac{1}{4} & (j = 4, \ldots, 9) \end{cases}$$

Упражнение 2. Покажите, что

$$\bar{p}_j^{(3)}(x_k, y_k) = \begin{cases} 1 & (k = j) \\ 0 & (k \neq j) \end{cases} \qquad (1 \leqslant j, \ k \leqslant 10).$$

Упражнение 3. Выразите $\bar{p}_j^{(3)}(x, y)$ $(j = 5, 6, 7, 8, 9)$ через $p_k(x, y)$ $(k = 1, 2, 3)$.

Вернемся теперь к общему случаю полного полинома m-го порядка (см. (4.1)). Этот полином имеет $\frac{1}{2}(m+1)(m+2)$ коэффициентов, которые могут быть выбраны так, чтобы полиномом интерполировал $U(x, y)$ по значениям в $\frac{1}{2}(m+1) \times (m+2)$ симметрично расположенных точках треугольника

$P_1P_2P_3$, координаты которых

$$\sum_{l=1}^{3} \frac{\beta_l x_l}{m}, \quad \sum_{l=1}^{3} \frac{\beta_l y_l}{m}, \tag{4.5}$$

где β_1, β_2, β_3 — целые числа, удовлетворяющие условиям $0 \leqslant \leqslant \beta_k \leqslant m$ $(k = 1, 2, 3)$ и $\beta_1 + \beta_2 + \beta_3 = m$. Эти точки включают три вершины треугольника $P_1P_2P_3$. Остальные точки получаются разбиением каждой стороны треугольника на m равных частей как точки пересечения прямых, параллельных сторонам треугольника и проходящих через точки разбиения. В результате получается разбиение треугольника на m^2 равных треугольников, чьи вершины суть $\frac{1}{2}(m+1)(m+2)$ точек, описанных в (4.5). Если обозначить через U_j значение $U(x, y)$ в такой точке, интерполирующий полином степени m можно выразить формулой

$$U(x, y) = \sum_{j=1}^{\frac{1}{2}(m+1)(m+2)} U_j p_j^{(m)}(x, y), \tag{4.6}$$

где суммирование осуществляется по всем точкам, а $p_j^{(m)}(x, y)$ — полиномиальная базисная функция степени m, принимающая значение 1 в точке, связанной с тройкой $(\beta_1, \beta_2, \beta_3)$, и значение 0 во всех остальных узлах. Формула (4.6) — интерполяционная формула *лагранжева типа.*

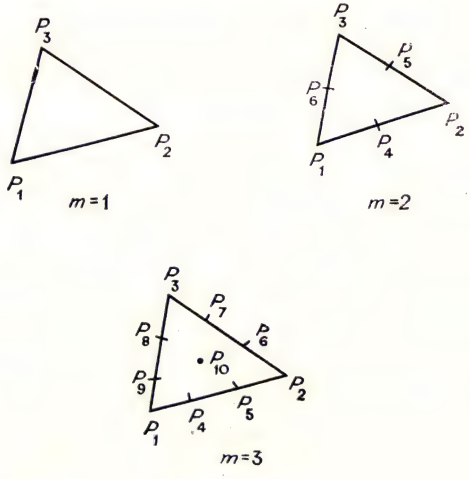


Рис. 11.

Стандартный треугольник

Из формулы (4.2) легко вывести, что

(I) $\sum_{j=1}^{3} p_j(x, y) = 1$,

(II) линейные уравнения $p_j(x, y) = 0$ $(j = 1, 2, 3)$ представляют стороны треугольника P_2P_3, P_3P_1, P_1P_2 соответственно и

(III) $p_j(x, y) = 1$ $(j = 1, 2, 3)$ в вершинах P_1, P_2, P_3 соответственно.

Иначе говоря, треугольник $P_1P_2P_3$ в плоскости (x, y) преобразуется в стандартный треугольник $\Pi_1\Pi_2\Pi_3$ в плоскости (p_1, p_2) по формулам (4.2), причем $\Pi_1 = (1, 0)$, $\Pi_2 = (0, 1)$, $\Pi_3 = (0, 0)$ (см. рис. 12). Обратное преобразование плоскости (p_1, p_2) в плоскость (x, y) выражается формулами

$$\begin{aligned} x &= x_3 + \xi_{13}p_1 + \xi_{23}p_2, \\ y &= y_3 + \eta_{13}p_1 + \eta_{23}p_2. \end{aligned} \tag{4.7}$$

Поскольку все треугольники треугольной сетки в плоскости (x, y) могут быть преобразованы в этот стандартный треугольник, очень удобно работать со стандартным треугольником и в нужный момент выражать результат в терминах конкретного треугольника плоскости (x, y) с помощью линейного преобразования (4.7). Эта процедура будет использоваться неоднократно в этой главе и в следующей, когда будут встречаться треугольные элементы.

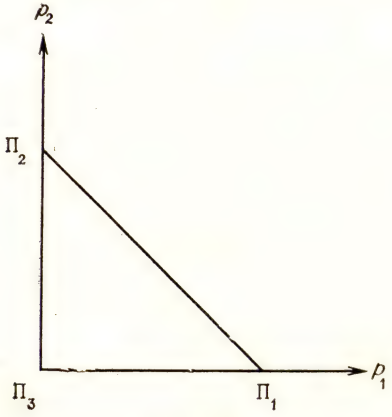


Рис. 12.

Геометрический метод построения базисных функций

Опишем теперь для треугольника в квадратичном и кубическом случаях простую геометрическую процедуру построения базисных функций, полученных из лагранжевой интерполяции. К примеру, в первом случае базисная функция $p_1^{(2)}(x, y)$ должна быть равной нулю в узлах p_j $(j = 2, 3, \ldots, 6)$ и единице в узле P_1. Прямая $p_1 = 0$ проходит через точки P_2, P_3 и P_5, а прямая линия $p_1 = \frac{1}{2}$ проходит через точки P_4, P_6, и таким образом получается базисная функция $p_1^{(2)}(x, y) = p_1(2p_1 - 1)$. Функции $p_1^{(2)}(x, y)$, $p_3^{(2)}(x, y)$ получаются аналогично. Базисная функция в узле, расположенном посредине стороны, скажем в P_4, получается из прямых $p_1 = 0$ и $p_2 = 0$, которые проходят через точки P_2, P_3, P_5 и P_1, P_3, P_6 соответственно. Требуемая базисная функция, нормированная в точке P_4, есть $p_4^{(2)}(x, y) = 4p_1p_2$; функции $p_5^{(2)}(x, y)$ и $p_6^{(2)}(x, y)$ получаются аналогично. В кубическом случае построения осуществляются аналогичным образом.

C^0-аппроксимирующие функции [1])

В общем случае функция интерполируется по значениям в $\frac{1}{2}(m + 1)(m + 2)$ точках в треугольнике. На стороне треугольника интерполирующая функция сводится к полиному степени m от переменной s, которая измеряется вдоль стороны треугольника. Этот полином интерполирует $U(x, y)$ по значениям в $(m + 1)$ точках на стороне треугольника и, значит, определен однозначно. Кроме того, при любой триангуляции каждая сторона (внутри области) принадлежит двум треугольникам. Если функция интерполирует в каждом треугольнике по $\frac{1}{2}(m + 1)(m + 2)$ симметрично расположенным точкам, она представляется единственным полиномом от s степени m на общей стороне. Это означает, что интерполирующая функция на полной треугольной сетке непрерывна на внутренних сторонах сетки и, таким образом, имеет на многоугольной области гладкость C^0.

(B) Эрмитова интерполяция

Как альтернативу интерполяции функции $U(x, y)$ на большом числе точек, симметрично расположенных в треугольни-

[1]) C^k-аппроксимирующими (или C^k-гладкими) называются такие аппроксимации, у которых все производные до порядка $k \geqslant 0$ включительно непрерывны во всей области. — *Прим. перев.*

ке, можно рассматривать интерполяцию $U(x, y)$ и некоторых ее производных по меньшему числу точек.

Класс полиномов, соответствующий этой задаче, содержит полные полиномы $G_\nu(x, y)$ нечетной степени $2\nu + 1$ ($\nu = 1, 2, 3, \ldots,$), которые определяются значениями

$$\begin{aligned} &D^i G_\nu(P_j) \quad (|i| \leqslant \nu;\ j = 1, 2, 3), \\ &D^i G_\nu(P_4) \quad (|i| \leqslant \nu - 1). \end{aligned} \tag{4.8}$$

Здесь P_1, P_2, P_3 — вершины треугольника, а P_4 — его центр тяжести; $i = (i_1, i_2)$, где i_1, i_2 — неотрицательные целые числа, $|i| = i_1 + i_2$ и $D^i G = \dfrac{\partial^{|i|} G}{\partial x^{i_1} \partial y^{i_2}}$.

Это пример *мультииндексных обозначений* производных. Первые два случая этого общего представления таковы:

(1) *Кубический случай* ($\nu = 1$). Здесь полный кубический полином имеет десять коэффициентов, которые определяются по значениям функции и ее первых частных производных в вершинах и по значению функции в центре тяжести треугольника, причем однозначно. Поэтому в этом случае мы можем записать полином $\Pi_3(x, y)$, данный в (4.4), в виде

$$G_1(x, y) = \sum_{j=1}^{4} U_j q_j^{(3)}(x, y) + \sum_{j=1}^{3} \left[\left(\frac{\partial U}{\partial x} \right)_j r_j^{(3)}(x, y) + \left(\frac{\partial U}{\partial y} \right)_j s_j^{(3)}(x, y) \right], \tag{4.9}$$

где

$$\begin{aligned} q_j^{(3)}(x, y) &= p_j (3p_j - 2p_j^2 - 7p_k p_l), \\ q_4^{(3)}(x, y) &= 27 p_1 p_2 p_3, \\ r_j^{(3)}(x, y) &= p_j [\xi_{jk} p_k (p_l - p_j) + \xi_{jl} p_l (p_k - p_j)], \end{aligned}$$

а $s_j^{(3)}(x, y)$ получается из $r_j^{(3)}(x, y)$ заменой ξ на η. В этих формулах (j, k, l) — любая циклическая перестановка $(1, 2, 3)$. Этот элемент широко используется в методе конечных элементов.

(2) *Случай полинома пятой степени* ($\nu = 2$). Полный полином пятой степени имеет 21 коэффициент; они однозначно определяются значениями функции и ее первых и вторых частных производных в вершинах и значениями функции и ее первых частных производных в центре тяжести. Этот элемент редко используется на практике и поэтому не рассматривается здесь подробно.

Упражнение 4. Покажите, что на стороне треугольника функция, заданная в (4.9), сводится к полиному степени 3 по переменной s, которая определяется вдоль стороны треугольника. Покажите также, что этот полином однозначно определяется значениями функции и ее первых частных производных в вершинах, являющихся концами стороны. Отсюда покажите, что интерполирующая функция этого элемента имеет гладкость C^0.

В случае кубической эрмитовой интерполяции можно исключить значение $U(x, y)$ в центре тяжести и еще точно интерполировать квадратичные полиномы. В этом случае делается замена

$$U_4 = \frac{1}{3}\sum_{j=1}^{3} U_j + \frac{1}{18}\sum^{(1)}\left[\left(\frac{\partial U}{\partial x}\right)_j (\xi_{kj} + \xi_{lj}) + \left(\frac{\partial U}{\partial y}\right)_j (\eta_{kj} + \eta_{lj})\right],$$

где $\sum^{(1)}$ обозначает суммирование по (j, k, l) для всех циклических перестановок (1, 2, 3).

Интерполирующий полином теперь имеет вид

$$G_1^*(x, y) = \sum_{j=1}^{3}\left[U_j q_j^{(3)*}(x, y) + \left(\frac{\partial U}{\partial x}\right)_j r_j^{(3)*}(x, y) + \left(\frac{\partial U}{\partial y}\right)_j s_j^{(3)*}(x, y)\right], \quad (4.10)$$

где

$$q_j^{(3)*}(x, y) = p_j(3p_j - 2p_j^2 + 2p_k p_l), \quad (4.10a)$$

$$r_j^{(3)*}(x, y) = p_j^2(p_l\xi_{lj} + p_k\xi_{kj}) + \frac{1}{2}p_j p_k p_l(\xi_{lj} + \xi_{kj}) \quad (4.10b)$$

и (j, k, l) — любая перестановка (1, 2, 3). Функции $s_j^{(3)*}$ получаются из (4.10b) заменой ξ на η.

Упражнение 5. Проверьте, что квадратичные функции точно интерполируются сокращенным кубическим интерполянтом.

Упражнение 6. Используя линейное преобразование по формулам (4.7), покажите, что (4.10) в переменных стандартного треугольника имеет вид

$$G_1^*(p_1, p_2) = \sum_{j=1}^{3}\left[U_j q_j^{(3)*} + \left(\frac{\partial U}{\partial p_1}\right)_j r_j^{(3)*} + \left(\frac{\partial U}{\partial p_2}\right)_j s_j^{(3)*}\right],$$

где

$$\frac{\partial U}{\partial p_j} = \xi_{j3}\frac{\partial U}{\partial x} + \eta_{j3}\frac{\partial U}{\partial y} \quad (j = 1, 2)$$

и функции $q_j^{(3)*}$, $r_j^{(3)*}$ и $s_j^{(3)*}$ $(j = 1, 2, 3)$ задаются формулами

$$q_j^{(3)*} = p_j^2(3 - 2p_j) + 2p_1p_2p_3,$$

$$r_j^{(3)*} = \begin{cases} p_1^2(p_1 - 1) - p_1p_2p_3 & (j = 1) \\ p_j^2p_1 + \frac{1}{2}p_1p_2p_3 & (j = 2, 3), \end{cases}$$

$$s_j^{(3)*} = \begin{cases} p_2^2(p_2 - 1) - p_1p_2p_3 & (j = 2) \\ p_j^2p_2 + \frac{1}{2}p_1p_2p_3 & (j = 1, 3). \end{cases}$$

Для доказательства можно воспользоваться соотношением

$$\xi_{kl}\eta_{lj} - \xi_{lj}\eta_{kl} = C_{jkl},$$

где (j, k, l) — любая перестановка $(1, 2, 3)$.

Упражнение 7. В частном случае $P_1 = (1, 0)$, $P_2 = (0, 1)$, $P_3 = (0, 0)$ найдите $G_1^*(x, y)$ из (4.10), а затем покажите, что производные от $G_1^*(x, y)$ по нормалям к сторонам описываются формулами

$$\left(\frac{\partial G_1^*}{\partial x}\right)_{P_2P_3} = y(1 - y)\left[2U_1 - \left(\frac{\partial U}{\partial x}\right)_1 + \frac{1}{2}\left(\frac{\partial U}{\partial y}\right)_1 + 2U_2 - \right.$$
$$\left. - \frac{1}{2}\left(\frac{\partial U}{\partial x}\right)_2 - \left(\frac{\partial U}{\partial y}\right)_2 - 4U_3 - \frac{1}{2}\left(\frac{\partial U}{\partial x}\right)_3 - \frac{3}{2}\left(\frac{\partial U}{\partial y}\right)_3\right] +$$
$$+ y\left(\frac{\partial U}{\partial x}\right)_2 + (1 - y)\left(\frac{\partial U}{\partial x}\right)_3,$$

$$\left(\frac{\partial G_1^*}{\partial y}\right)_{P_3P_1} = x(1 - x)\left[2U_1 - \left(\frac{\partial U}{\partial x}\right)_1 - \frac{1}{2}\left(\frac{\partial U}{\partial y}\right)_1 + 2U_2 + \right.$$
$$\left. + \frac{1}{2}\left(\frac{\partial U}{\partial x}\right)_2 - \left(\frac{\partial U}{\partial y}\right)_2 - 4U_3 - \frac{3}{2}\left(\frac{\partial U}{\partial x}\right)_3 - \frac{1}{2}\left(\frac{\partial U}{\partial y}\right)_3\right] +$$
$$+ x\left(\frac{\partial U}{\partial y}\right)_1 + (1 - x)\left(\frac{\partial U}{\partial y}\right)_3,$$

$$\left(\frac{\partial G_1^*}{\partial X}\right)_{P_1P_2} = \frac{1}{4}(1 - Y)^2\left[U_1 - \frac{1}{2}\left(\frac{\partial U}{\partial X}\right)_1 + \frac{1}{2}\left(\frac{\partial U}{\partial Y}\right)_1 + U_2 - \right.$$
$$\left. - \frac{1}{2}\left(\frac{\partial U}{\partial X}\right)_2 - \frac{1}{2}\left(\frac{\partial U}{\partial Y}\right)_2 - 2U_3 - \left(\frac{\partial U}{\partial X_3}\right)\right] +$$
$$+ \frac{1}{2}(1 - Y)\left(\frac{\partial U}{\partial X}\right)_1 + \frac{1}{2}(1 + Y)\left(\frac{\partial U}{\partial X}\right)_2,$$

где

$$X = x + y, \quad Y = y - x, \quad \frac{\partial U}{\partial X} = \frac{1}{2}\left(\frac{\partial U}{\partial x} + \frac{\partial U}{\partial y}\right),$$
$$\frac{\partial U}{\partial Y} = \frac{1}{2}\left(-\frac{\partial U}{\partial x} + \frac{\partial U}{\partial y}\right).$$

Трикубическая интерполяция

Биркгоф (1971) предложил треугольный элемент, включающий двенадцатипараметрическое семейство всех полиномов четвертой степени, которые кубичны вдоль любой прямой, параллельной любой стороне треугольника. Относительно стандартного треугольника такое семейство выражается формулой

$$U(p, q) = \sum_{j+k \leqslant 4} \alpha_{jk} p_1^j p_2^k, \tag{4.11}$$

причем

$$\alpha_{31} + \alpha_{13} = \alpha_{22}.$$

Полином (4.11) называется *трикубическим*. Этот полином однозначно определяется по значениям U, $\partial U/\partial p_1$, $\partial U/\partial p_2$ и $\frac{\partial^2 U}{\partial r\, \partial s}$ в каждой вершине. Здесь $\frac{\partial^2 U}{\partial r\, \partial s}$ — смешанная производная, вычисляемая в каждой вершине по направлениям r и s, параллельным смежным сторонам. Эту единственную интерполирующую функцию можно представить в виде

$$\sum_{j=1}^{3}\left[\left\{\sum_{|i| \leqslant 1} D^i U_j \varphi_j^i(p_1, p_2)\right\} + \left(\frac{\partial^2 U}{\partial r\, \partial s}\right)_j \hat{\varphi}_j(p_1, p_2)\right], \tag{4.12}$$

где коэффициентами при φ_j^i и $\hat{\varphi}_j$ служат значения соответствующих производных от U в вершине Π_j стандартного треугольника (рис. 12). Функции φ_j^i и $\hat{\varphi}_j$ имеют вид

$$\varphi_j^{(0,0)} = p_j^2(3 - 2p_j + 6p_k p_l),$$
$$\varphi_j^{(1,0)} = \begin{cases} p_1^2(p_1 - 1 - 4p_2 p_3) & (j = 1) \\ p_2^2 p_1(1 + 2p_3) & (j = 2) \\ p_3^2 p_1(1 + 2p_2) & (j = 3), \end{cases}$$
$$\varphi_j^{(0,1)} = \begin{cases} p_1^2 p_2(1 + 2p_3) & (j = 1) \\ p_2^2 p_2(p_2 - 1 - 4p_1 p_3) & (j = 2) \\ p_3^2 p_2(1 + 2p_1) & (j = 3), \end{cases}$$
$$\hat{\Phi}_j = 2p_j^2 p_k p_l,$$

где (j, k, l) — любая перестановка (1, 2, 3). Заметим, что здесь D^i представляют производные в плоскости (p_1, p_2). Смешанные производные $(\partial^2 U/\partial r \partial s)_j$ $(j = 1, 2, 3)$ суть просто $-\frac{\partial^2 U}{\partial p_1 \partial Q}$, $\frac{\partial^2 U}{\partial p_2 \partial Q}$, $\frac{\partial^2 U}{\partial p_1 \partial p_2}$ соответственно, где $Q = p_2 - p_1$. Единственная интерполирующая трикубическая функция для конкретного треугольника на плоскости (x, y) получается из (4.12) с помощью линейного преобразования (4.2), т. е. подстановкой вместо p_1, p_2, p_3 их выражений из (4.2).

Упражнение 8. Покажите, что трикубические полиномы дают аппроксимирующую функцию гладкости C^0 (т. е. просто непрерывную) на сетке треугольных элементов.

(C) C^1-аппроксимирующие функции

Рассматриваемый здесь треугольный элемент использует полное семейство полиномов пятой степени. На плоскости стандартного треугольника полный полином пятой степени имеет вид

$$U(p_1, p_2) = \sum_{j+k \leqslant 5} \alpha_{jk} p_1^j p_2^k. \tag{4.13}$$

Коэффициенты α_{jk} можно выразить через $D^i U(P_j)$ $(|i| \leqslant 2;$ $j = 1, 2, 3)$ и $\partial U/\partial n(P_j)$ $(j = 4, 5, 6)$, где P_j $(j = 1, 2, 3)$ — вершины, а P_j $(j = 4, 5, 6)$ — середины сторон. Функция $U(p_1, p_2)$, заданная в (4.13), сводится к полиному пятой степени по s вдоль каждой стороны треугольника, определенному однозначно 6 граничными условиями, а именно значениями U, $\partial U/\partial s$, $\frac{\partial^2 U}{\partial s^2}$, на концах каждой стороны. Нормальная производная на каждой стороне, $\frac{\partial U}{\partial n}$, где n суть p_2, $p_1 + p_2$, p_1 соответственно, есть полином четвертой степени по s; он определяется однозначно значениями $\frac{\partial U}{\partial n}$ и $\frac{\partial^2 U}{\partial n \partial s}$ на концах каждой стороны и значением $\frac{\partial U}{\partial n}$ в середине стороны. Отсюда видно, что полные полиномы пятой степени дают аппроксимирующую функцию на сетке треугольных элементов, непрерывную вместе со своими первыми производными на всей области. Про такую функцию говорят, что она имеет на области гладкость C^1.

На самом деле параметры, соответствующие нормальным производным в серединах сторон, можно исключить без потери C^1-гладкости на треугольной сетке. Это можно сделать, если предположить кубическое поведение нормальной производной вдоль каждой стороны, а это эквивалентно требова-

нию, что в (4.13)

$$\alpha_{41} = \alpha_{14},$$

$$5\alpha_{50} + \alpha_{32} + \alpha_{23} + 5\alpha_{05} = 0.$$

Таким путем от (4.13) мы приходим к семейству полиномов, зависящему от 18 параметров, и однозначно определенная интерполирующая функция получается по формуле

$$U(p_1, p_2) = \sum_{j=1}^{3} \sum_{i \leqslant 2} D^i U_j \varphi_j^i(p_1, p_2), \tag{4.14}$$

где

$$\varphi_j^{(0,0)} = \begin{cases} p_j^2(10p_j - 15p_j^2 + 6p_j^2 + 15p_k^2 p_l) & (j = 1, 2) \\ p_j^2(10p_j - 15p_j^2 + 6p_j^3 + 30p_k p_l(p_k + p_l)) & (j = 3) \end{cases}$$

и (j, k, l) — циклические перестановки $(1, 2, 3)$,

$$\varphi_1^{(1,0)} = p_1^2\left(-4p_1 + 7p_1^2 - 3p_1^3 - \frac{15}{2}p_2^2 p_3\right),$$

$$\varphi_2^{(1,0)} = p_1 p_2^2\left(3 - 2p_2 - \frac{3}{2}p_1 - \frac{3}{2}p_1^2 + \frac{3}{2}p_1 p_2\right),$$

$$\varphi_3^{(1,0)} = p_1 p_3^2(3 - 2p_3 - 3p_1^2 + 6p_1 p_2),$$

$$\varphi_1^{(0,1)} = p_1^2 p_2\left(3 - 2p_1 - \frac{3}{2}p_2 - \frac{3}{2}p_2^2 + \frac{3}{2}p_1 p_2\right),$$

$$\varphi_2^{(0,1)} = p_2^2\left(-4p_2 + 7p_2^2 - 3p_2^3 - \frac{15}{2}p_1^2 p_3\right),$$

$$\varphi_3^{(0,1)} = p_2 p_3^2(3 - 2p_3 - 3p_2^2 + 6p_1 p_2),$$

$$\varphi_1^{(1,1)} = p_1^2 p_2\left(-1 + p_1 + \frac{1}{2}p_2 + \frac{1}{2}p_2^2 - \frac{1}{2}p_1 p_2\right),$$

$$\varphi_2^{(1,1)} = p_1 p_2^2\left(-1 + \frac{1}{2}p_1 + p_2 + \frac{1}{2}p_1^2 - \frac{1}{2}p_1 p_2\right),$$

$$\varphi_3^{(1,1)} = p_1 p_2 p_3^2,$$

$$\varphi_1^{(2,0)} = p_1^2\left(\frac{1}{2}p_1(1 - p_1)^2 + \frac{5}{4}p_2^2 p_3\right),$$

$$\varphi_2^{(2,0)} = \frac{1}{4}p_1^2 p_2^2 p_3 + \frac{1}{2}p_1^2 p_2^3,$$

$$\varphi_3^{(2,0)} = \frac{1}{2}p_1^2 p_3^2(1 - p_1 + 2p_2),$$

$$\varphi_1^{(0,2)} = \frac{1}{4}p_1^2 p_2^2 p_3 + \frac{1}{2}p_1^3 p_2^2,$$

$$\varphi_2^{(0,2)} = p_2^2\left(\frac{1}{2}p_2(1 - p_2)^2 + \frac{5}{4}p_1^2 p_3\right),$$

$$\varphi_3^{(0,2)} = \frac{1}{2}p_2^2 p_3^2(1 + 2p_1 - p_2).$$

Корректирующие функции

В другом методе получения C^1-аппроксимирующей функции на треугольной сетке берется C^0-аппроксимирующая функция из (4.10) и к ней добавляются корректирующие члены, которые повышают гладкость функции до C^1. Эти корректирующие функции должны обращаться в нуль на периметре треугольника и сводить нормальную производную функции вдоль сторон треугольника от квадратичной к линейной функции, разумеется, без потери непрерывности функции и ее первых производных внутри треугольного элемента.

Одно семейство корректирующих функций, широко используемых на практике (Зенкевич, 1967, с. 113), имеет вид

$$A_j(p_1, p_2, p_3) = \frac{p_j p_k^2 p_l^2}{(1-p_k)(1-p_l)}, \tag{4.15}$$

где (j, k, l) — перестановка (1, 2, 3). Фактически Дюпюи и Гёль (1970) показали, что C^1-гладкость получается, если к правым частям (4.10a) и (4.10b) добавить

$$\delta q_k = 2\left[-A_j + \left(2 - 3\frac{L_l}{L_k}\cos\theta_j\right)A_k + \left(2 - 3\frac{L_k}{L_l}\cos\theta_j\right)A_l\right], \tag{4.15a}$$

$$\delta r_j = \frac{1}{2}\left[(\xi_{jk} + \xi_{jl})A_j + \left(3\xi_{lj} + 5\xi_{jk} + 6\eta_{lj}\frac{L_l}{L_k}\sin\theta_j\right)A_k + \right.$$
$$\left. + \left(3\xi_{kj} + 5\xi_{jl} + 6\eta_{jk}\frac{L_k}{L_l}\sin\theta_j\right)A_l\right], \tag{4.15b}$$

где (j, k, l) — любая циклическая перестановка (1, 2, 3), θ_j — угол треугольника у вершины P_j, а L_j — длина стороны, лежащей против вершины P_j; δs_j получается из (4.15b) заменой ξ и η на η и ξ соответственно.

Упражнение 9. Покажите, что в случае $P_1 = (1, 0)$, $P_2 = (0, 1)$ и $P_3 = (0, 0)$ дополнительные члены имеют вид

$$\delta q_3 = -2\,\delta q_2 = -2\,\delta q_1 = 4\,\delta r_1 = 4\,\delta s_2 = 4\bar{A} + 4\bar{B} - 2\bar{C},$$

$$\delta r_2 = -\delta s_1 = \frac{1}{2}\bar{A} - \frac{1}{2}\bar{B},$$

$$\delta r_3 = \frac{1}{2}(\bar{A} + 3\bar{B} - \bar{C}),$$

$$\delta s_3 = \frac{1}{2}(3\bar{A} + \bar{B} - \bar{C}),$$

где

$$\bar{A}=\frac{xy^2(1-x-y)^2}{(1-y)(x+y)},\quad \bar{B}=\frac{x^2y(1-x-y)^2}{(1-x)(x+y)},$$
$$\bar{C}=\frac{x^2y^2(1-x-y)}{(1-x)(1-y)}.$$

Отсюда, используя результат упражнения 7, покажите, что нормальные производные линейны вдоль каждой стороны треугольника.

Другие корректирующие функции, указанные Клафом и Точером (1965), получаются так. Каждый треугольник разбивается на три треугольника с общей вершиной в центре тяжести. Корректирующие функции выражаются формулой

$$A_j(p_1, p_2, p_3)=\begin{cases} p_1p_2p_3-\frac{1}{6}p_j^2(3-5p_j) & \text{на } T_j \\ \frac{1}{6}p_k^2(3p_l-p_k) & \text{на } T_k \\ \frac{1}{6}p_l^2(3p_k-p_l) & \text{на } T_l, \end{cases} \tag{4.16}$$

где T_j — треугольник напротив вершины P_j и т. д., а (j, k, l) — любая перестановка (1, 2, 3). Добавки δq_j, δr_j и δs_j получаются из (4.15a) и (4.15b) с A_j из (4.16).

Упражнение 10. Для треугольника упражнения 9 найдите корректирующие функции $\bar{A}$, $\bar{B}$, $\bar{C}$ из (4.16) и покажите, что на стороне, описываемой уравнением $x=0$,

$$\frac{\partial\bar{A}}{\partial x}=-6y(1-y),\quad \frac{\partial\bar{B}}{\partial x}=\frac{\partial\bar{C}}{\partial x}=0,$$

на стороне с $y=0$

$$\frac{\partial\bar{B}}{\partial y}=-6x(1-x),\quad \frac{\partial\bar{A}}{\partial y}=\frac{\partial\bar{C}}{\partial y}=0,$$

и на стороне, точки которой удовлетворяют уравнению $x+y=1$,

$$\frac{\partial\bar{C}}{\partial X}=\frac{3}{2}(1-Y^2),\ \frac{\partial\bar{A}}{\partial X}=\frac{\partial\bar{C}}{\partial X}=0,$$

где $X=x+y$ и $Y=y-x$. Покажите, что отсюда следует линейное изменение нормальных производных от $q_l+\delta q_l$, $r_j+\delta r_j$ и $s_j+\delta s_j$ вдоль сторон. (*Указание.* Используйте результат упражнения 7.)

Упражнение 11. Пусть функции $A_j(x, y)$ имеют вид

$$\Delta_{jk}=\sum_{l+m+n=3}\beta_{lmn}p_1^lp_2^mp_3^n$$

на малых треугольниках T_k. Проверьте, что

(I) в вершинах $D^i A_j = 0$ ($|i| \leqslant 1$) и
(II) на сторонах

$$\frac{\partial A_j}{\partial n} = \begin{cases} 0 & (p_j \neq 0) \\ p_k p_l & (p_j = 0) \end{cases}$$

тогда и только тогда, когда

(I) $\Delta_{jj} = p_1 p_2 p_3 + p_j^2 (\alpha_{jj} p_j + \alpha_{jk} p_k + \alpha_{jl} p_l)$ и

(II) $\Delta_{jk} = p_j^2 (\alpha_{kj} p_j + \alpha_{kk} p_k + \alpha_{kl} p_l)$

для любых констант α_{jj}, α_{jk} и т. д., где (j, k, l) — любая перестановка (1, 2, 3).

После этого докажите, что функции A_j, заданные в (4.16), имеют C^1-гладкость на треугольнике $P_1P_2P_3$, если приравнять Δ_{jj} и т. д. и их производные вдоль внутренних линий раздела.

Последний набор корректирующих функций таков:

$$A_j(p_1, p_2, p_3) = \frac{p_j^2 p_k}{1 - p_l},$$

где (j, k, l) — циклическая перестановка (1, 2, 3). Биркгоф и Менсфилд (1974) использовали его для добавления к трикубическому полиному с целью получить C^1-гладкость на треугольной сетке. Это C^1-гладкое семейство, зависящее от 15 параметров, легко выразить через значения U, $\partial U/\partial p_1$, $\partial U/\partial p_2$ в вершинах и $\partial U/\partial n$, $\partial^2 U/\partial s \partial n$ в серединах сторон, где $\partial/\partial s$ и $\partial/\partial n$ обозначают частные производные по направлениям сторон и нормалей соответственно. Айронс (1969) использовал некоторые корректирующие функции для добавления к полному полиному четвертой степени и таким образом получил семейство, зависящее от 18 параметров, которое также имеет C^1-гладкость на треугольной сетке.

4.2. Прямоугольник

Области типа прямоугольника, т. е. области со сторонами, параллельными осям x и y, возникают во многих задачах физики и техники. Следовательно, прямоугольный элемент имеет большое значение и в этом параграфе для него строятся базисные функции.

Бикубические эрмитовы функции

В параграфе 1.1 мы использовали билинейные функции от x и y для построения базисных функций (1.17), каждая из которых равна нулю вне области, составленной из четырех

прямоугольных элементов. На всей прямоугольной области кусочно-билинейные аппроксимирующие функции имеют C^0-гладкость. В этом пункте мы рассмотрим бикубические полиномы

$$H_3(x,\ y) = \sum_{j=0}^{3} \sum_{k=0}^{3} \alpha_{jk} x^j y^k \tag{4.17}$$

на единичном квадрате $0 \leqslant x,\ y \leqslant 1$. Коэффициенты α_{jk} $(0 \leqslant j,\ k \leqslant 3)$ можно однозначно найти по значениям H_3, $\frac{\partial H_3}{\partial x}$, $\frac{\partial H_3}{\partial y}$, $\frac{\partial^2 H_3}{\partial x\, \partial y}$ в четырех вершинах, так что

$$H_3(x,\ y) = \sum_{0 \leqslant j,\ k,\ l,\ m \leqslant 1} (D^{(l,\ m)}U)_{jk}\, \psi_j^{(l)}(x)\, \psi_k^{(m)}(y), \tag{4.18}$$

где

$$\psi_0^{(0)}(t) = (1-t)^2(1+2t),$$
$$\psi_0^{(1)}(t) = (1-t)^2 t,$$
$$\psi_1^{(0)}(t) = t^2(3-2t),$$
$$\psi_1^{(1)}(t) = t^2(t-1).$$

Нижний индекс jk указывает значение в вершине с $x = j$, $y = k$.

Нетрудно увидеть, как можно изменить (4.18) для получения требуемой эрмитовой бикубической интерполирующей функции на прямоугольном элементе исходной области типа прямоугольника. Аппроксимирующая функция на всей области получается затем аналогично выводу (1.6). На этот раз каждому узлу прямоугольной области соответствуют *четыре* базисные функции. Для внутреннего узла, т. е. узла не на границе прямоугольной области, каждая базисная функция имеет своим носителем четыре прямугольных элемента. Для узла на границе, но не в углу, базисная функция имеет носителем два элемента, а для узла в углу — один прямоугольный элемент. Кусочно-бикубическая аппроксимирующая функция на всей прямоугольной области оказывается $C^{1,\,1}$-гладкой [1]).

Упражнение 12. Используя (4.18), покажите, что базисные функции для $D^{(l,\ m)}U$ $(0 \leqslant l,\ m \leqslant 1)$ в узле $(0, 0)$ *единичной ячейки* получаются в *форме тензорного произведения*

$$\varphi^{(l,\ m)}(x,\ y) = \varphi^{(l)}(x)\, \varphi^{(m)}(y),$$

[1]) $C^{j,\,k}$-гладкость $u(x, y)$ означает непрерывность всех производных $D^{(l,\ m)}$ и $(0 \leqslant l \leqslant j,\ 0 \leqslant m \leqslant k)$.

где

$$\varphi^{(0)}(t) = \begin{cases} (1-t)^2(1+2t) & (0 \leqslant t \leqslant 1) \\ (1+t)^2(1-2t) & (-1 \leqslant t \leqslant 0), \end{cases}$$

$$\varphi^{(1)}(t) = \begin{cases} (1-t)^2 t & (0 \leqslant t \leqslant 1) \\ (1+t)^2 t & (-1 \leqslant t \leqslant 0). \end{cases}$$

Интересный прямоугольный элемент был описан Пауэллом (1973). Прямоугольник разбивается диагоналями на четыре треугольника, и предполагается, что на каждом треугольнике аппроксимирующая функция задается полной квадратичной по x и y функцией. Коэффициенты этих квадратичных функций можно выбрать так, что получится C^1-гладкость на прямоугольной сетке.

Бикубические сплайны

В разд. 1.1 было показано, что одномерный кубический сплайн с локальным носителем длины $4h$ имеет вид (1.15). Фактически этот сплайн, впервые предложенный Шёнбергом (1969), часто записывается как

$$M(x) = \frac{1}{6h}\delta^4\left(\frac{x}{h} - j\right)_+^3, \tag{4.19}$$

где δ — обычный оператор центральной разности, а константа $\frac{1}{6h}$ выбрана так, что

$$\int_{-\infty}^{+\infty} M(x)\,dx = 1.$$

Константа $\frac{1}{4}$ в (1.15) была выбрана с тем, чтобы

$$B_I(j) = 1.$$

В областях типа прямоугольника, разбитых на прямоугольные элементы, рассмотрим сплайны Шёнберга $M(x)$ из (4.19) и

$$M(y) = \frac{1}{6h}\delta^4\left(\frac{y}{h} - k\right)_+^3.$$

Тензорное произведение $M(x)$ и $M(y)$ дает функцию с носителем из 16 прямоугольных элементов. Она оказывается базисной функцией в узле $x = jh$, $y = kh$ для кубических сплайнов, если только узел не лежит на границе области и не примыкает к ней. Для граничных и примыкающих к границе уз-

лов нужно строить специальные базисные функции, если, конечно, решаемая задача не имеет естественных граничных условий (гл. 3, стр. 54). Полная аппроксимирующая функция в этом случае является $C^{2,2}$-гладкой.

Возможно, заслуживает упоминания то, что в маловероятном случае $C^{4,4}$-гладкости аппроксимирующей функции на прямоугольной области, если только она потребуется, могут быть использованы, подобно кубическим сплайнам, сплайны Шёнберга пятой степени

$$M(x) = \frac{1}{5!h}\delta^6\left(\frac{x}{h} - j\right)_+^5$$

Их тензорное произведение имеет носитель из 36 прямоугольных элементов, что делает эти сплайны довольно неудобными для работы.

В заключение этого короткого параграфа о прямоугольном элементе следует указать на то, что размер носителя сплайна увеличивается с ростом порядка сплайна, тогда как носитель эрмитовой функции остается всегда состоящим из четырех элементов, независимо от ее порядка, в то же время сплайны имеют гладкость $C^{2(\nu-1),\,2(\nu-1)}$ против $C^{\nu-1,\,\nu-1}$ для эрмитовых функций при степени полиномов $2\nu - 1$.

4.3. Четырехугольник

Можно подумать, что четырехугольник является лучшей формой ячейки, чем треугольник, поскольку сетка в целом упрощается. Например, треугольная сетка всегда может быть упрощена объединением треугольников попарно в четырехугольники. К сожалению, однако, невозможно найти полином от x и y, который бы сводился к произвольной линейной форме вдоль четырех сторон общего четырехугольника, и поэтому не ясно, как можно построить кусочно-полиномиальную функцию от x и y, которая имела бы C^0-гладкость на четырехугольной сетке.

Лемма 4.1. *Пусть* $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$ *и* $\mathcal{P}_4$ *— точки в трехмерном пространстве, такие, что* $\mathcal{P}_j = (x_j, y_j, z_j)$ $(j = 1, 2, 3, 4)$. *Тогда плоскость, проходящая через три точки* $\mathcal{P}_j$, $\mathcal{P}_k$ *и* $\mathcal{P}_l$ $(j \neq k \neq l)$, *описывается уравнением*

$$\Pi_{jkl} = 0,$$

где $\Pi_{jkl} = -zC_{jkl} + z_jD_{kl} - z_kD_{jl} + z_lD_{jk}$, *а* C_{jkl}, D_{kl} *и т. д. определены в разд. 4.1. Кроме того, поверхность*

$$\alpha\Pi_{klm}\Pi_{jkm} - \beta\Pi_{jlm}\Pi_{jkl} = 0, \qquad (4.20)$$

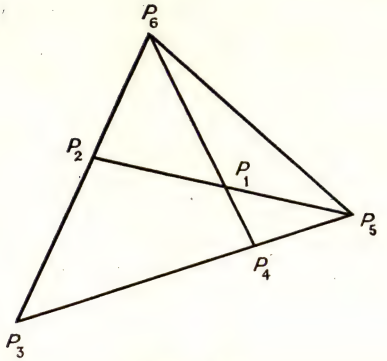


Рис. 13.

где (j, k, l, m) *— любая циклическая перестановка* $(1, 2, 3, 4)$ *проходит через четыре точки* $\mathscr{P}_1$, $\mathscr{P}_2$, $\mathscr{P}_3$ *и* $\mathscr{P}_4$ *и содержит прямые линии* $\mathscr{P}_1\mathscr{P}_2$, $\mathscr{P}_2\mathscr{P}_3$, $\mathscr{P}_3\mathscr{P}_4$, $\mathscr{P}_4\mathscr{P}_1$ *при любых значениях* α *и* β.

Из этой леммы ясно, что такие поверхности как (4.20), которые проходят через точки $\mathscr{P}_j = (x_j, y_j, f_j)$ $(j = 1, 2, 3, 4)$, можно использовать для определения функций $f(x, y)$ на четырехугольнике $P_1P_2P_3P_4$, где $P_j = (x_j, y_j)$, таких, что (I) $f(x_j, y_j) = f_j$ и (II) $f(x, y)$ изменяется линейно вдоль сторон P_1P_2, P_2P_3, P_3P_4, P_4P_1 четырехугольника.

Таким образом, можно взять $\mathscr{P}_j = (x_j, y_j, 1)$ и $\mathscr{P}_k = (x_k, y_k, 0)$ $(k \neq j)$ и определить базисную функцию $\varphi_j(x, y)$, такую, что

$$\varphi_j(x_k, y_k) = \begin{cases} 1 & (j = k) \\ 0 & (j \neq k) \end{cases} \qquad (1 \leqslant j, k \leqslant 4). \tag{4.21}$$

Упражнение 13. Пусть P_5 — точка пересечения прямых P_1P_2 и P_3P_4, а P_6 — точка пересечения прямых P_2P_3 и P_4P_1 (рис. 13). Докажите, что прямая линия P_5P_6 $(D_{56} = 0)$ такова, что

$$\text{(I)} \quad \frac{D_{23}}{C_{123}} + \frac{D_{34}}{C_{134}} - \frac{D_{24}}{C_{124}} = \frac{D_{56}}{C_{156}} \tag{4.22a}$$

и что существует такая константа $\lambda \neq 0$, для которой

$$\text{(II)} \quad \begin{aligned} &C_{134}C_{234}D_{12} + C_{123}C_{124}D_{34} = \\ &= C_{123}C_{234}D_{14} + C_{134}C_{124}D_{23} = \lambda D_{56}. \end{aligned} \tag{4.22b}$$

Отсюда покажите, что если $\alpha C_{klm}C_{jkm} = \beta C_{jlm}C_{jkl}$, то базисные функции, определенные в (4.21), могут быть записаны

как

$$\frac{D_{kl}}{C_{jkl}}\frac{D_{lm}}{C_{jlm}}\left(\frac{D_{56}}{C_{j56}}\right)^{-1},$$

где P_kP_l и P_lP_m — стороны четырехугольника, не содержащие угол P_j, и где (j, k, l, m) — некоторая перестановка $(1, 2, 3, 4)$. (Используйте формулы

$$\text{(I)}\quad C_{jkl} - C_{klm} + C_{jlm} - C_{jkm} = 0, \tag{4.23a}$$

$$\text{(II)}\quad C_{jkl} - D_{kl} + D_{jl} - D_{ik} = 0, \tag{4.23b}$$

где (j, k, l, m) — любая перестановка $(1, 2, 3, 4)$.)

Изопараметрические координаты

Билинейная аппроксимация

Наиболее общий метод использования четырехугольных элементов состоит в применении точечного преобразования четырехугольника в единичный квадрат и в использовании так называемой изопараметрической аппроксимации (Айронс, 1966; Зенкевич, 1975). Другими словами, угловые точки $P_j = (x_j, y_j)$ $(j = 1, 2, 3, 4)$ четырехугольника преобразуются в четыре точки $(1, 1)$, $(0, 1)$, $(0, 0)$ и $(1, 0)$ (в плоскости (p, q)). Стандартное преобразование есть

$$t = pqt_1 + (1-p)\,qt_2 + (1-p)(1-q)\,t_3 + p(1-q)\,t_4 \quad (t = x,\ y), \tag{4.24}$$

которое можно записать как

$$t = \sum_{j=1}^{4} \varphi_j^{(1)}(p, q)\,t_j \quad (t = x,\ y). \tag{4.25}$$

Изопараметрическая аппроксимация получается, если определить аппроксимацию вида (4.25), а именно

$$U(p, q) = \sum_{j=1}^{4} \varphi_j^{(1)}(p, q)\,U_j. \tag{4.26}$$

Упражнение 14. Покажите, что преобразование, обратное (4.25), может быть записано как

$$(C_{234} + C_{134})\,p^2 + (D_{34} + D_{12} - C_{123} - C_{234})\,p + D_{23} = 0 \tag{4.27a}$$

и

$$(C_{234} + C_{123})\,q^2 + (D_{23} + D_{41} - C_{134} - C_{234})\,q + D_{34} = 0. \tag{4.27b}$$

Далее, используя (4.23a) и (4.23b), покажите, что функция p, определенная в (4.27a), эквивалентна поверхности вида

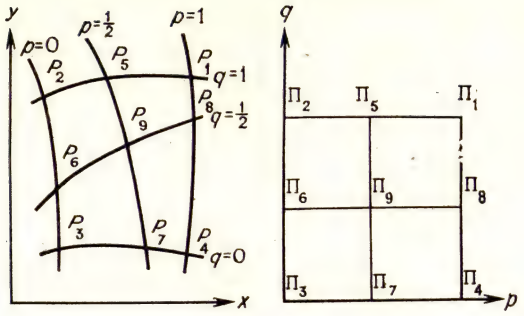


Рис. 14.

(4.20), проходящей через точки $\mathscr{P}_j = (x_j, y_j, f_j)$ с $f_2 = f_3 = 0$ и $f_1 = f_4 = 1$, если $\alpha = \beta$, а q, определенная в (4.27b), эквивалентна такой же поверхности с $f_3 = f_4 = 0$, $f_1 = f_2 = 1$ и $\alpha = \beta$.

Упражнение 15. Если требуется преобразовать четырехугольник $P_1P_2P_3P_4$ в единичный квадрат, то выбор $\alpha = \beta$ не является единственно возможным. Покажите, что если $\mathscr{P}_j$ ($j = 1, 2, 3, 4$) взяты как в упражнении 14, и $\alpha C_{klm}C_{jkm} = \beta C_{jlm}C_{jkl}$, то новые координаты p и q можно определить так:

$$p = \frac{D_{23}}{C_{123}}\frac{D_{34}}{C_{134}}\left(\frac{D_{56}}{C_{156}}\right)^{-1} + \frac{D_{23}}{C_{234}}\frac{D_{12}}{C_{124}}\left(\frac{D_{56}}{C_{456}}\right)^{-1},$$

$$q = \frac{D_{23}}{C_{123}}\frac{D_{34}}{C_{134}}\left(\frac{D_{56}}{C_{156}}\right)^{-1} - \frac{D_{34}}{C_{234}}\frac{D_{14}}{C_{124}}\left(\frac{D_{56}}{C_{256}}\right)^{-1}.$$

Упражнение 16. Покажите, что якобиан J преобразования (4.25) можно записать как

$$J = (1-p)\,C_{123} + (1-q)\,C_{134} + (p+q-1)\,C_{124},$$

и докажите, что $J > 0$ для $0 \leqslant p, q \leqslant 1$.

Если билинейные полиномы, использованные в преобразовании (4.25), заменить на полиномы более высокой степени, можно ввести дополнительные точки, определяющие преобразование, и одновременно распространить изопараметрические аппроксимации на криволинейные четырехугольники.

Биквадратичная аппроксимация

Теперь добавим к четырем точкам $P_j = (x_j, y_j)$ ($j = 1, 2, 3, 4$), которые соответствуют углам единичного квадрата в (p, q)-плоскости, точки P_j ($j = 5, \ldots, 9$), которые соответ-

ствуют серединам сторон $\left(\frac{1}{2}, 1\right)$, $\left(0, \frac{1}{2}\right)$, $\left(\frac{1}{2}, 0\right)$ и $\left(1, \frac{1}{2}\right)$ и центру $\left(\frac{1}{2}, \frac{1}{2}\right)$ соответственно (рис. 14). Биквадратичное преобразование определяется как

$$t = \sum_{j=1}^{9} \varphi_j^{(2)}(p, q)\, t_j \quad (t = x, y), \tag{4.28}$$

где $\varphi_1^{(2)} = p(2p-1)\,q(2q-1)$ и аналогично определяются $\varphi_2^{(2)}\varphi_3^{(2)}$ и $\varphi_4^{(2)}$, $\varphi_5^{(2)} = 4(1-p)\,pq(2q-1)$ и аналогично определяются $\varphi_6^{(2)}\varphi_7^{(2)}$ и $\varphi_8^{(2)}$, а $\varphi_9^{(2)} = 16p(1-p)\,q(1-q)$.
Изопараметрическая аппроксимация тогда определяется формулой

$$U(p, q) = \sum_{j=1}^{9} \varphi_j^{(2)}(p, q)\, U_j. \tag{4.29}$$

С другой стороны, если преобразование определено по (4.25), а аппроксимация — по (4.29), то это — пример *субпараметрической аппроксимации.*

Стороны четырехугольника можно сделать прямыми подходящей расстановкой узлов P_5, P_6, P_7 и P_8 на сторонах. В частности, если они являются серединами соответствующих сторон, а P_9 — центром тяжести четырехугольника, формулы (4.28) сводятся к (4.25).

Внутренний узел P_9 можно исключить аналогично тому, как исключались внутренние узлы из аппроксимаций, основанных на треугольных элементах, используя линейное соотношение

$$U_9 = \frac{1}{2}\sum_{j=5}^{8} U_j - \frac{1}{4}\sum_{j=1}^{4} U_j.$$

Это дает функцию, которая все еще интерполирует квадратичные по p и q функции точно, но *не имеет члена* с p^2q^2. Ее можно записать как

$$U(p, q) = \sum_{j=1}^{8} \varphi_j^{(2)*}(p, q)\, U_j, \tag{4.30}$$

где $\varphi_1^{(2)*} = pq(2p + 2q - 3)$ и аналогично $\varphi_j^{(2)*}$ $(j = 2, 3, 4)$, $\varphi_5^{(2)*} = 4pq(1-p)$ и аналогично $\varphi_j^{(2)*}$ $(j = 6, 7, 8)$. Для изопараметрической аппроксимации с восемью узлами преобразование (x, y) в (p, q) будет также определяться по (4.30) заменой U поочередно на x и y. Этот четырехугольник с восемью узлами

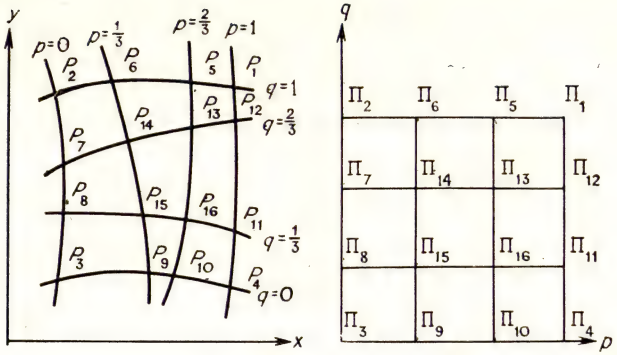


Рис. 15.

был использован Джорданом (1970) в качестве элемента при решении задач, включающих плоское напряжение или деформацию.

Бикубическая аппроксимация

Полная бикубическая аппроксимация включает четыре внутренних узла (см. рис. 15) в дополнение к четырем угловым и восьми боковым узлам (по два на каждой стороне). Внутренние узлы могут быть исключены для получения аппроксимации, в которой нет членов с p^2q^2, p^3q^2, p^2q^3 и p^3q^3. Ее можно записать в виде

$$U(p, q) = \sum_{j=1}^{12} \varphi_j^{(3)*}(p, q)\, U_j,$$

где $\varphi_1^{(3)*} = \frac{9}{2}\left(p^2 + q^2 - p - q + \frac{2}{9}\right)pq$ и аналогично $\varphi_j^{(3)*}$ $(j = 2, 3, 4)$, $\varphi_5^{(3)*} = \frac{9}{2}pq(1-p)(3p-1)$ и аналогично $\varphi_j^{(3)*}$ $(j = 6, \ldots, 12)$.

4.4. Тетраэдр [1])

Лагранжева интерполяция

Полный полином степени

$$\Pi_m(x, y, z) = \sum_{i+k+l=0}^{m} \alpha_{ikl} x^i y^k z^l$$

[1]) Сравните с разд. 4.1.

можно использовать для интерполяции функции $U(x, y, z)$ по $\frac{(m+1)(m+2)(m+3)}{6}$ симметрично расположенным в тетраэдре узлам. Первые три случая этого общего представления для тетраэдра $P_1P_2P_3P_4$ таковы:

(1) *Линейный случай* ($m = 1$). Полином задается формулой

$$\Pi_1(x, y, z) = \sum_{j=1}^{4} U_j p_j^{(1)}(x, y, z),$$

где U_j ($j = 1, 2, 3, 4$) — значения $U(x, y, z)$ в вершинах P_j и

$$p_j^{(1)}(x, y, z) = \frac{1}{\Gamma_{jkln}}(E_{kln} - A_{kln}x + B_{kln}y - C_{kln}z),$$

где

$$\Gamma_{jkln} = \det\begin{bmatrix} 1 & x_j & y_j & z_j \\ 1 & x_k & y_k & z_k \\ 1 & x_l & y_l & z_l \\ 1 & x_n & y_n & z_n \end{bmatrix}, \quad A_{kln} = \det\begin{bmatrix} 1 & y_k & z_k \\ 1 & y_l & z_l \\ 1 & y_n & z_n \end{bmatrix}$$

и т. д., а (j, k, l, n) — любая перестановка $(1, 2, 3, 4)$. Заметим, что $\frac{1}{6}|\Gamma_{jkln}|$ — объем тетраэдра $P_1P_2P_3P_4$.

(2) *Квадратичный случай* ($m = 2$). Теперь полином таков:

$$\Pi_2(x, y, z) = \sum_{j=1}^{10} U_j p_j^{(2)}(x, y, z),$$

где U_j ($j = 1, 2, 3, 4$) — значения $U(x, y, z)$ в вершинах p_j, а U_j ($j = 5, \ldots, 10$) — значения в серединах ребер. Выражение $p_j^{(2)}$ (через $p_j^{(1)}$) имеет тот же вид, что и для треугольника (см. параграф 4.1).

(3) *Кубический случай* ($m = 3$). Здесь полином имеет вид

$$\Pi_3(x, y, z) = \sum_{j=1}^{20} U_j p_j^{(3)}(x, y, z),$$

где U_j ($j = 1, 2, 3, 4$) — как и выше, U_j ($j = 5, \ldots, 15$) — значения $U(x, y, z)$ в точках трисекции ребер, а U_j ($j = 16, \ldots, 20$) — значения в центрах тяжести граней. Формулы для $p_j^{(3)}$ (через $p_j^{(1)}$) имеют тот же вид, что и для треугольника (см. разд. 4.1).

Эрмитова интерполяция

Кубическую аппроксимацию $U(x, y, z)$ на тетраэдре можно выписать подобно (4.9)—(4.10b) как

$$G_1(x, y, z) =$$

$$= \sum_{j=1}^{4} \left[U_j q_j^{(3)} + \left(\frac{\partial U}{\partial x}\right)_j r_j^{(3)} + \left(\frac{\partial U}{\partial y}\right)_j s_j^{(3)} + \left(\frac{\partial U}{\partial z}\right)_j t_j^{(3)} + \bar{U}_j \bar{q}_j^{(3)} \right], \quad (4.31)$$

где $\bar{U}_j$ $(j = 1, \ldots, 4)$ — значения $U(x, y, z)$ в центрах тяжести граней, противолежащих вершинам P_j, и где

$$q_j^{(3)}(x, y, z) = p_j (3p_j - 2p_j^2 - 7(p_k p_l + p_l p_n + p_n p_k)), \quad (4.32a)$$

$$\bar{q}_j^{(3)}(x, y, z) = 27 p_k p_l p_n,$$

и

$$r_j^{(3)}(x, y, z) = p_j [(\xi_{jk} p_k (p_l + p_n - p_j) +$$

$$+ \xi_{jl} p_l (p_k + p_n - p_j) + \xi_{jn} p_n (p_k + p_l - p_j)]; \quad (4.32b)$$

$s_j^{(3)}(x, y, z)$ получается из $r_j^{(3)}(x, y, z)$ заменой ξ на η, а $t_j^{(3)}(x, y, z)$ — заменой ξ на ζ, где $\zeta_{jk} = z_j - z_k$. Как и для треугольника, мы писали p_j вместо $p_j^{(1)}$ $(j = 1, 2, 3, 4)$ для упрощения формул.

Упражнение 17. Проверьте, что можно исключить значения $U(x, y, z)$ в центрах тяжести граней и все еще точно интерполировать квадратичные функции, если сделать замену

$$\bar{U}_j = \frac{1}{3} \sum_{k \neq j} U_k + \frac{1}{18} \sum^{(1)} \left[\left(\frac{\partial U}{\partial x}\right)_k (\xi_{nk} + \xi_{lk}) + \right.$$

$$\left. + \left(\frac{\partial U}{\partial y}\right)_k (\eta_{nk} + \eta_{lk}) + \left(\frac{\partial U}{\partial z}\right)_k (\zeta_{nk} + \zeta_{lk}) \right],$$

где $\sum^{(1)}$ означает суммирование (при фиксированном j) по (j, k, l, n) для всех возможных четных перестановок $(1, 2, 3, 4)$. Покажите, что интерполирующий полином становится таким:

$$G_1^*(x, y, z) = \sum_{j=1}^{4} \left[U_j q_j^{(3)*} + \left(\frac{\partial U}{\partial x}\right)_j r_j^{(3)*} + \left(\frac{\partial U}{\partial y}\right)_j s_j^{(3)*} + \left(\frac{\partial U}{\partial z}\right)_j t_j^{(3)*} \right],$$

где

$$q_j^{(3)*}(x, y, z) = p_j (3p_j - 2p_j^2 + 2(p_k p_l + p_k p_n + p_l p_n)),$$

$$r_j^{(3)*}(x, y, z) = p_j^2 (\xi_{kj} p_k + \xi_{lj} p_l + \xi_{nj} p_n) + \frac{1}{2} p_j \{ p_k p_l (\xi_{kj} + \xi_{lk}) +$$

$$+ p_k p_n (\xi_{kj} + \xi_{nj}) + p_l p_n (\xi_{lj} + \xi_{nj}) \}$$

и т. д.

4.5. Шестигранник [1])

Вообще говоря, в случае трехмерного пространства элемент с шестью четырехугольными гранями лучше тетраэдра. Локальные изопараметрические координаты (p, q, r) можно ввести как и для четырехугольных элементов в двумерном случае. Преобразование координат имеет вид

$$t = \sum_{j=1}^{8} \varphi_j^{(1)}(p, q, r)\, t_j \quad (t = x, y, z), \tag{4.33}$$

где $\varphi_1^{(1)} = pqr$, $\varphi_2^{(1)} = (1-p)\,qr$ и т. д., причем вершины пронумерованы как на рис. 16. При таком преобразовании произвольный шестигранник переходит в единичный куб в (p, q, r)-пространстве. Затем по формуле

$$U(p, q, r) = \sum_{j=1}^{8} \varphi_j^{(1)}(p, q, r)\, U_j \tag{4.34}$$

определяется изопараметрическая аппроксимация. Можно получить и триквадратичную и трикубическую аппроксимации, если ввести дополнительные точки на ребрах и гранях, а также ряд внутренних точек. Точки на гранях и внутренние точки можно исключить подобно тому, как это было сделано в двумерном случае при исключении внутренних точек четырехугольника.

Упражнение 18. Вычислите базисные функции $\varphi_j^{(2)}(p, q, r)$ $(j = 1, \ldots, 27)$ для триквадратичной изопараметрической аппроксимации на шестиграннике. Затем проверьте, что центры тяжести граней и самого шестигранника можно исключить и получить аппроксимацию вида

$$U(p, q, r) = \sum_{j=1}^{20} \varphi_j^{(2)*}(p, q, r)\, U_j, \tag{4.35}$$

которая точно интерполирует квадратичные функции, но не имеет членов с p^2q^2, q^2r^2, r^2p^2, p^2q^2r, p^2qr^2, pq^2r^2, $p^2q^2r^2$ и для которой $\varphi_1^{(2)*} = pqr\,(2p + 2q + 2r - 5)$ и аналогично $\varphi^{(2)*}$ $(j = 2, \ldots, 8)$, $\varphi_9^{(2)*} = 4pqr\,(1-p)$ и аналогично $\varphi_j^{(2)*}$ $(j = 10, \ldots, 20)$.

Упражнение 19. Проверьте, что трикубическая аппроксимация на шестиграннике требует 64 узлов: восемь вершин, по две точки на каждом ребре, по четыре — на каждой грани и восемь точек внутри шестигранника. Затем проверьте, что узлы на гранях и внутренние узлы могут быть исключены для получения аппроксимации, которая включает члены

[1]) Сравните с четырехугольником в параграфе 4.3.

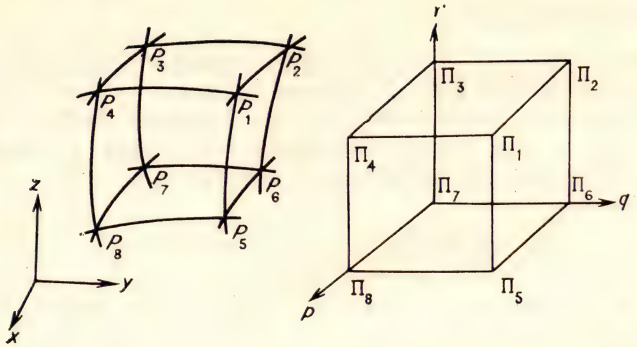


Рис. 16.

с $p^j q^k r^l (j + k + l \leqslant 4)$, а также члены пятой степени p^3qr, pq^3r и pqr^3, так что

$$U(p, q, r) = \sum_{j=1}^{32} \varphi_j^{(3)*}(p, q, r) U_j, \qquad (4.36)$$

где $\varphi_1^{(3)*} = \frac{9}{2} pqr \left[p^2 + q^2 + r^2 - (p + q + r) + \frac{2}{9} \right]$ и аналогично $\varphi_j^{(3)*}$ $(j = 2, \ldots, 8)$, а $\varphi_9^{(3)*} = \frac{9}{2} pqr (1 - p)(3p - 1)$ и аналогично $\varphi_j^{(3)*}$ $(j = 10, \ldots, 32)$.

4.6. Криволинейные границы

До сих пор базисные функции строились в основном для сеток с прямыми сторонами ячеек. В реальных двумерных и трехмерных задачах, однако, границы и поверхности раздела часто криволинейны. Цель этого параграфа заключается в получении базисных функций для сеток, составленных из элементов с криволинейными сторонами (двумерный случай). или с криволинейными поверхностями (трехмерный случай). Криволинейный элемент появился при расчете сооружений у Эргатодиса, Айронса и Зенкевича (1968), и библиографические разъяснения по этому вопросу можно найти у Зенкевича (1975). В двумерном случае, если граница области является ломаной линией, элементов с прямыми сторонами, обычно треугольников или четырехугольников, вполне достаточно. Однако если некоторая часть границы (или линии раздела материалов) изогнута, желательны элементы по крайней мере с одной криволинейной стороной.

Вначале мы рассмотрим треугольный элемент с двумя прямыми и одной криволинейной сторонами. С помощью

этого элемента и треугольников с прямыми сторонами можно адекватно решить большинство плоских задач с криволинейными границами и линиями раздела.

(A) Треугольники с одной криволинейной стороной

Рассмотрим треугольник $P_1P_2P_3$ в (x, y)-плоскости с прямыми сторонами P_2P_3 и P_3P_1, которые описываются уравнениями $l(x, y) = 0$ и $m(x, y) = 0$ соответственно. Криволинейная сторона, проходящая через точки P_1 и P_2, описывается уравнением $F(x, y) = 0$. Функции $l(x, y)$, $m(x, y)$ и $F(x, y)$ нормированы так, что

$$l(x_1, y_1) = m(x_2, y_2) = F(x_3, y_3) = 1.$$

Преобразование плоскости (x, y) в плоскость (l, m) выглядит так:

$$l = \frac{1}{C_{123}}(\tau_{23} + \eta_{23}x - \xi_{23}y), \tag{4.37a}$$

$$m = \frac{1}{C_{123}}(\tau_{31} + \tau_{31}x - \xi_{31}y). \tag{4.37b}$$

В плоскости (l, m) получается треугольник $P_1'P_2'P_3'$, где $P_1' = (1, 0)$, $P_2' = (0, 1)$ и $P_3' = (0, 0)$, а криволинейная сторона $P_1'P_2'$ выражается уравнением

$$F(x(l, m), y(l, m)) = f(l, m) = 0.$$

Мы используем здесь l и m вместо $p_1^{(1)}$ и $p_2^{(1)}$ (ср. с (4.2)), потому что треугольник $P_1'P_2'P_3'$ с одной криволинейной стороной не является стандартным треугольником.

Теперь мы опишем один тип лагранжевой аппроксимации на криволинейном треугольнике $P_1'P_2'P_3'$, которая удовлетворяет следующим условиям:

(I) Линейные полиномы интерполируются точно, т. е.

$$\sum_i \varphi_i = 1; \quad \sum_i l_i\varphi_i = l; \quad \sum_i m_i\varphi_i = m, \tag{4.38}$$

где $\varphi_i(l, m)$ — базисная функция, связанная с i-узлом. Так как преобразование (x, y) в (l, m) линейно, соотношения (4.38) означают, что линейные полиномы в плоскости (x, y) на треугольнике $P_1P_2P_3$ также интерполируются точно.

(II) Базисная функция $\varphi_3(l, m)$, соответствующая P_3', равна тождественно нулю на криволинейной стороне $P_2'P_1'$.

(III) Получающаяся кусочно-гладкая функция, определенная на сетке треугольников, каждый из которых имеет самое большее одну криволинейную сторону, непрерывна.

Нетрудно увидеть, что для выполнения (I) и (II) на треугольнике $P_1'P_2'P_3'$ необходимо иметь по крайней мере четыре узла. В простейшем случае берутся три вершины и дополнительная точка $P_4' = (l_4, m_4)$ на криволинейной стороне.

Вначале мы построим базисную функцию φ_3, которая удовлетворяет (II), а затем используем (4.38) для построения φ_1, φ_2 и φ_4. Применим геометрические рассмотрения, подобные тем, которые были привлечены к получению базисных функций для четырехугольника, и поэтому рассмотрим семейство поверхностей $z(l, m) = 0$, которое пересекает плоскость (l, m) по кривой $f(l, m) = 0$ и задается уравнением

$$z(\alpha z + \beta l + \gamma m + \delta) + f(l, m) = 0. \tag{4.39}$$

Если мы наложим на z условия

$$\begin{aligned} &(1)\quad z = 1 && (l = m = 0),\\ &(2)\quad z = 1 - l && (m = 0)\\ &(3)\quad z = 1 - m && (l = 0), \end{aligned}$$

то можно определить β, γ, δ с помощью одной произвольной константы α, так что (4.39) перейдет в

$$\alpha z^2 + \left[\alpha(l+m-1) + 1 - \frac{f(l, 0)}{1-l} - \frac{f(0, m)}{1-m}\right] z + f(l, m) = 0. \tag{4.40}$$

Теперь мы полагаем $\varphi_3 = z$, после чего остальные базисные функции определяются из (4.38) как

$$\begin{aligned} \varphi_1 &= \frac{(1-m_4)\,l + l_4 m - l_4}{1 - l_4 - m_4} + \frac{l_4}{1 - l_4 - m_4}\,\varphi_3,\\ \varphi_2 &= \frac{m_4 l + (1-l_4)\,m - m_4}{1 - l_4 - m_4} + \frac{m_4}{1 - l_4 - m_4}\,\varphi_3,\\ \varphi_4 &= \frac{1 - l - m}{1 - l_4 - m_4} - \frac{1}{1 - l_4 - m_4}\,\varphi_3. \end{aligned} \tag{4.41}$$

Упражнение 20. Проверьте, что если в (4.40) $\alpha = 0$, базисная функция φ_3 имеет вид

$$\varphi_3 = \frac{f(l, m)}{f(0, m)/(1-m) + f(l, 0)/(1-l) - 1}, \tag{4.42}$$

а если криволинейная сторона является коническим сечением

$$f(l, m) = al^2 + blm + cm^2 - (1+a)\,l - (1+c)\,m + 1 = 0,$$

то

$$\varphi_3 = \frac{al^2 + blm + cm^2 - (1+a)\,l - (1+c)\,m + 1}{1 - al - cm}, \tag{4.43}$$

и мы приходим к рациональным базисным функциям Уачспресса (1971, 1973, 1974 и 1975).

Упражнение 21. Если криволинейная сторона — отрезок гиперболы

$$f(l, m) = blm - l - m + 1 = 0,$$

покажите, что базисные функции φ_i $(i = 1, 2, 3, 4)$ суть полиномы, если $\alpha = 0$.

Замечание. Кусочно-гиперболические дуги можно использовать для аппроксимации криволинейных границ или поверхностей раздела и получить все еще полиномиальные базисные функции.

(B) Изопараметрические координаты [1])

Квадратичная аппроксимация на стандартном треугольнике

Часто треугольные элементы с криволинейными границами преобразуют в стандартный треугольник, а потом используют изопараметрические аппроксимации. Проиллюстрируем этот подход вначале на примере общего криволинейного треугольника, изображенного на рис. 17. Преобразование (x, y)-плоскости в (p, q)-плоскость задается формулой

$$t = \sum_{j=1}^{6} p_j^{(2)}(p, q)\, t_j \quad (t = x, y), \tag{4.44}$$

где квадратичные базисные функции $p_j^{(2)}$ $(j = 1, \ldots, 6)$ определяются как в разд. 4.1, только p_1 и p_2 заменены на p и q соответственно. Тогда изопараметрическая аппроксимация получается аналогично, т. е.

$$U(p, q) = \sum_{j=1}^{6} p_j^{(2)} U_j. \tag{4.45}$$

Если мы рассмотрим частный случай треугольника с двумя прямыми сторонами и одной криволинейной, так что $t_5 = \frac{t_2 + t_3}{2}$, $t_6 = \frac{t_3 + t_1}{2}$ $(t = x, y)$, то преобразование (4.44) сводится к

$$l = \alpha pq + p, \tag{4.46a}$$

$$m = \beta pq + q, \tag{4.46b}$$

где $\alpha = 2(2l_4 - 1)$ и $\beta = 2(2m_4 - 1)$. Обратное преобразование выглядит так:

$$\beta p^2 + (\alpha m - \beta l + 1)\, p - l = 0, \tag{4.47a}$$

$$\alpha q^2 + (\beta l - \alpha m + 1)\, q - m = 0 \tag{4.47b}$$

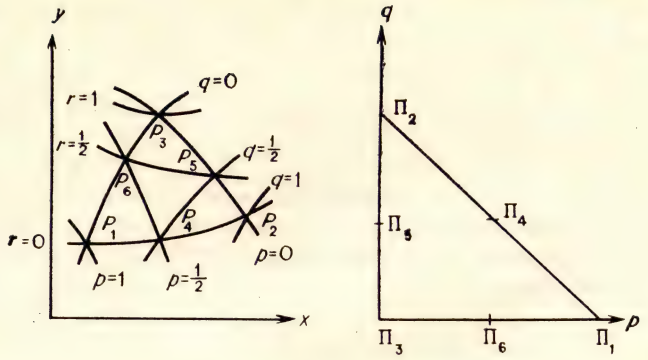


Рис. 17.

Упражнение 22. Покажите, что если $l_4 = m_4 = R$, то из (4.47a) и (4.47b) следует, что $r(= 1 - p - q)$ удовлетворяет уравнению

$$r^2 - \frac{4R-1}{2R-1}\, r + \left[\frac{2R - l - m}{2R-1} - (l-m)^2\right] = 0 \qquad (4.47c)$$

и, следовательно, криволинейная сторона $f(l, m) = 0$ заменяется кривой $r = 0$, описываемой уравнением

$$1 - \frac{l+m}{2R} - \frac{2R-1}{2R}(l-m)^2 = 0, \qquad (4.48)$$

которая оказывается параболой. Покажите далее, что если $\alpha = \frac{2R-1}{2R}$ и $f(l, m)$ определена по (4.48), то (4.40) переходит в (4.47c) с заменой z на r. Объясните эту связь между изопараметрической аппроксимацией и прямыми методами работы с криволинейными границами.

Упражнение 23. Покажите, что вообще криволинейная сторона $r = 0$ задается уравнением

$$(\beta l - \alpha m)^2 + (\alpha + \beta + \alpha\beta - \beta^2)\, l + (\alpha + \beta + \alpha\beta - \alpha^2)\, m - \\ - (\alpha + \beta + \alpha\beta) = 0,$$

где α и β заданы выше.

Запрещенные элементы

Одним из неприятных моментов использования изопараметрических координат для работы с криволинейными элементами является случай обращения в нуль якобиана $J(= 1 + \beta p + \alpha q)$ преобразования, заданного в (4.46a) и

(4.46b). Этот якобиан положителен для всех p, q, таких, что $0 \leqslant p,\ q,\ p+q \leqslant 1$, если только точка (l_4, m_4) лежит в области $l,\ m > \frac{1}{4}$, как показано на рис. 18. При других положениях точки (l_4, m_4) в положительном квадранте (l, m)-плоскости, включая и линии $l = \frac{1}{4}$ и $m = \frac{1}{4}$, якобиан либо равен нулю, либо отрицателен для некоторых значений (p, q) (Джордан, 1970), и поэтому в этих «запрещенных» случаях изопараметрические координаты, вообще говоря, нельзя использовать для работы с криволинейными элементами (исключения из этого правила приводятся в разделе 7.4(F)). Причина этого заключается в том, что результаты, вычисленные с помощью изопараметрических координат (p, q), нельзя перенести обратно на (l, m)-плоскость, поскольку вследствие обращения в нуль якобиана где-то на элементе обратного преобразования нет.

Вместо использования изопараметрических координат мы можем работать прямо с l и m, применяя (4.40) и (4.41). Если, как в упражнении 22, искривленная сторона определяется из (4.48) и $\alpha = (2R-1)/2R$, уравнение (4.40) будет иметь вещественные корни, если только

$$F(l, m; R) = \left(l + m + \frac{1}{2(2R+1)}\right)^2 - 4lm \geqslant 0. \qquad (4.49)$$

Нетрудно заметить, что при фиксированном R функция $F(l, m; R)$ не имеет максимума или минимума внутри элемента или где-либо на (l, m)-плоскости. Следовательно, наименьшее значение $F(l, m; R)$ достигается на границе элемента при всех значениях R. В самом деле, из (4.48) сле-

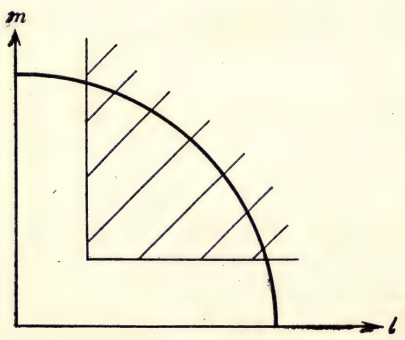


Рис. 18.

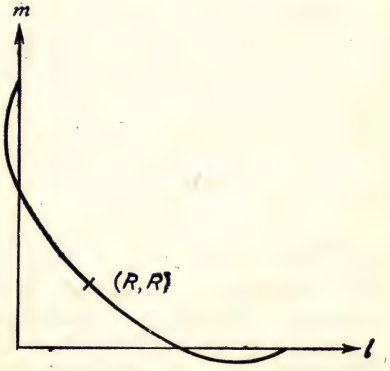


Рис. 19.

дует, что

$$F(l, m; R) = \left(\frac{4R-1}{4R-2}\right)^2 \geqslant 0$$

для всех R на криволинейной стороне, и, конечно, F всегда положительно на сторонах $l=0$ и $m=0$ в силу (4.49). Таким образом, условие (4.49) выполняется для всех R и для всех точек (l, m) элемента.

При использовании в этом примере изопараметрических координат значения R в области $0 < R < \frac{1}{4}$ использовать запрещается. Может показаться, что при использовании (4.40) таких ограничений нет. Однако с помощью простых геометрических рассмотрений можно показать, что при $0 < R < \frac{1}{4}$ криволинейная граница пересекает оси l и m в точках между началом координат и единичными точками осей (см. рис. 19).

Упражнение 24. Для $0 < R < \frac{1}{4}$ найдите точки пересечения криволинейной стороны с осями l и m между началом координат и единичными точками и покажите, что при $R \to \frac{1}{4}$ эти точки стремятся к единичным точкам на соответствующих осях.

Упражнение 25. Покажите, что квадратное уравнение (4.40) имеет вещественные корни при любых значениях α, когда точка (l, m) лежит на границе треугольника с двумя прямыми сторонами и одной криволинейной стороной.

Кубическая аппроксимация на стандартном треугольнике

Однозначно определенный кубический интерполяционный полином на плоскости (p, q) может быть записан как

$$U(p, q) = \sum_{j=1}^{10} p_j^{(3)} U_j, \tag{4.50}$$

где базисные функции $p_j^{(3)}$ $(j = 1, \ldots, 10)$ описаны в разд. 4.1. Мы получим изопараметрическую аппроксимацию, если используем (4.50) для определения точечного преобразования (x, y) (или (l, m)) в (p, q) заменой U на x и y (или l и m) (см. рис. 20). В частном случае треугольника с двумя прямыми сторонами, у которого P_6, P_7 и P_8, P_9 — точки трисек-

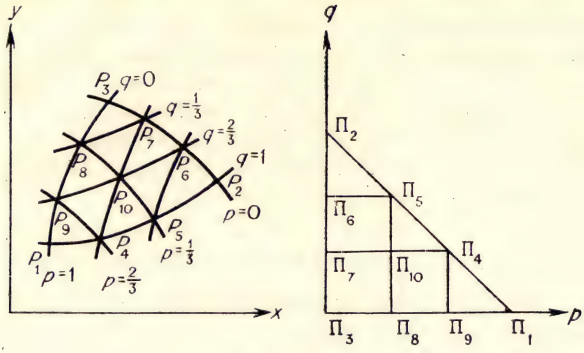


Рис. 20.

ции сторон P_2P_3 и P_3P_1 соответственно, мы получаем формулы

$$l = p + \frac{9}{2}\,pq\,(6l_{10} - l_4 - l_5 - 1) + \frac{27}{2}\,p^2q\,(l_4 - 2l_{10}) + \\ + \frac{27}{2}\,pq^2\left(l_5 - 2l_{10} + \frac{1}{3}\right), \qquad (4.51)$$

$$m = q + \frac{9}{2}\,pq\,(6m_{10} - m_4 - m_5 - 1) + \\ + \frac{27}{2}\,p^2q\left(m_4 - 2m_{10} + \frac{1}{3}\right) + \frac{27}{2}\,pq^2\,(m_5 - 2m_{10}),$$

где $P'_j = (l_j, m_j)$ $(j = 4, 5)$ — точки на криволинейной стороне, а (l_{10}, m_{10}) — внутри треугольника. Из (4.51) следует, что преобразованная кривая, проходящая через точки P'_1, P'_4, P'_5 и P'_2, оказывается *кубической*.

Упражнение 26. Покажите, что если $l_4 = l_5 + \frac{1}{3}$ и $m_4 = m_5 - \frac{1}{3}$, то кубическая кривая вырождается в *единственную* параболу, проходящую через точки $(1, 0)$, (l_4, m_4), (l_5, m_5) и $(0, 1)$. Если к тому же $l_4 = 2l_{10}$ и $m_5 = 2m_{10}$, то докажите, что формулы преобразования (4.51) сводятся к

$$l = p + 9\left(l_{10} - \frac{1}{3}\right)pq,$$

$$m = q + 9\left(m_{10} - \frac{1}{3}\right)pq$$

и что уравнение кривой из упражнения 23 при $4l_4 = 9l_{10} - 1$ и $4m_4 = 9m_{10} - 1$ задает параболу данного упражнения.

В статье Маклеода и Митчелла (1975) построены примеры разнообразных криволинейных сторон и получены соответствующие дуги парабол. Во всех примерах было обнаружено, что параболы близки к первоначальным кривым. Важно подчеркнуть то, что изопараметрические элементы вообще чрезвычайно чувствительны к деформации основной треугольной формы.

(C) Четырехугольник с одной криволинейной стороной

В разд. 4.3 для биквадратичной аппроксимации на четырехугольнике была дана формула (4.30). В частном случае четырехугольника с тремя прямыми и одной криволинейной сторонами (рис. 21), в котором P_6, P_7 и P_8 — середины прямых сторон P_2P_3, P_3P_4 и P_4P_1 соответственно, формулы точечного преобразования, соответствующие (4.30), сводятся к

$$\begin{aligned} l &= p + (-1 - l_1 + 4l_5)\,pq + (2l_1 - 4l_5)\,p^2q, \\ m &= q + (-3 - m_1 + 4m_5)\,pq + (2 + 2m_1 - 4m_5)\,p^2q \end{aligned} \tag{4.52}$$

для изопараметрической аппроксимации, где

$$l = \frac{1}{C_{234}}(\tau_{23} + \eta_{23}x - \xi_{23}y),$$

$$m = \frac{1}{C_{234}}(\tau_{34} + \eta_{34}x - \xi_{34}y).$$

Координату q можно исключить из (4.52) и получить

$$Tp^3 + [Z - Tl + Ym]\,p^2 + [1 + Xm - Zl]\,p - l = 0.$$

Кривая $q = 1$ задается уравнением

$$[Tl - Y(1 - m)]^2 + (T + XT - YZ)\,[Zl + (1 + X)(1 - m)] = 0. \tag{4.53}$$

Здесь $X = -1 - l_1 + 4l_5$, $Y = 2l_1 - 4l_5$, $Z = -3 - m_1 + 4m_5$ и $T = 2 + 2m_1 - 4m_5$. Эта кривая, конечно, — *парабола*. Следовательно, если используются изопараметрические координаты, как определено формулами точечного преобразования, соответствующими (4.30), *криволинейная сторона заменяется дугой параболы, уравнением которой является* (4.53).

Упражнение 27. Для случая четырехугольника с одной криволинейной стороной и с дополнительными узлами в точках трисекции всех прямых сторон (рис. 22) покажите, что

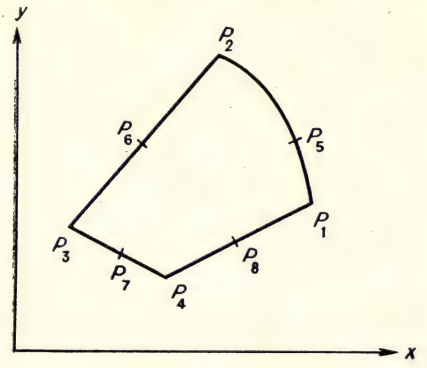


Рис. 21.

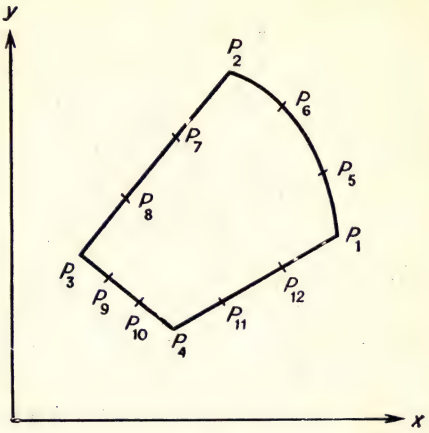


Рис. 22.

формулы точечного преобразования сводятся к

$$l = p + \frac{9}{2}\left[\left(\frac{2}{9}l_1 - l_5 + 2l_6 - \frac{2}{9}\right)pq - (l_1 - 4l_5 + 5l_6)\,p^2q + (l_1 - 3l_5 + 3l_6)\,p^3q\right],$$

$$m = q + \frac{9}{2}\left[\left(\frac{2}{9}m_1 - m_5 + 2m_6 - \frac{11}{9}\right)pq - (m_1 - 4m_5 + 5m_6 - 2)\,p^2q + (m_1 - 3m_5 + 3m_6 - 1)\,p^3q\right].$$

Затем покажите, что изопараметрическая координата q может быть исключена, чтобы получить уравнение четвертой степени по p, что кривая $q = 1$ — четвертого порядка по l и m, а кривые постоянных p — прямые линии.

(D) Тетраэдр с криволинейными гранями

Рассмотрим тетраэдр с четырьмя криволинейными гранями, у которого на каждом из шести искривленных ребер взят один промежуточный узел (см. рис. 23(а)). Стандартный тетраэдр в (p, q, r)-пространстве можно отобразить в (x, y, z)-пространство так, что его вершины и промежуточные узлы перейдут в соответствующие вершины и узлы произвольного наперед заданного искривленного тетраэдра. Это осуществляется с помощью точечного преобразования

$$t = p(2p-1)t_1 + q(2q-1)t_2 + r(2r-1)t_3 + s(2s-1)t_4 + 4pst_5 + 4pqt_8 + 4qst_6 + 4rqt_9 + 4prt_{10} + 4rst_7, \quad (4.54)$$

где $p + q + r + s = 1$, $t = x, y, z$. Для изопараметрического элемента интерполирующая функция также задается в виде

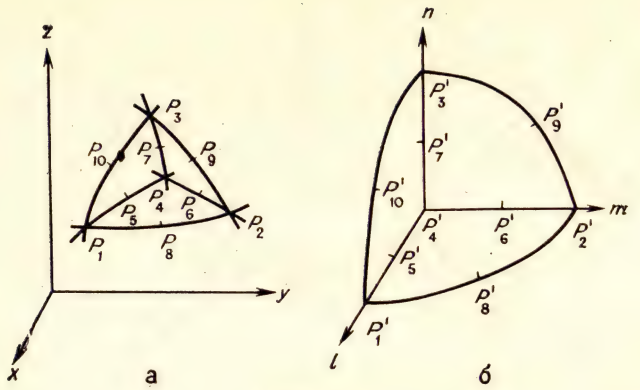


Рис. 23.

(4.54) с заменой t на U. Так как большинство ограниченных областей в трехмерном пространстве можно приближенно разбить [1]) на тетраэдральные элементы либо с четырьмя плоскими гранями, либо с одной искривленной и тремя плоскими гранями, мы остановимся на последнем тетраэдре. Если плоские грани $P'_2P'_3P'_4$, $P'_3P'_1P'_4$ и $P'_1P'_2P'_4$ лежат в плоскостях $l=0$, $m=0$ и $n=0$ соответственно, а вершины P'_1, P'_2, P'_3 на осях соответствуют $l=1$, $m=1$, $n=1$, то мы приходим к рис. 23(b), где P'_5, P'_6 и P'_7 — середины соответствующих прямых ребер, а $P'_8=(R_1, R_1, 0)$, $P'_9=(0, R_2, R_2)$ и $P'_{10}=(R_3, 0, R_3)$. Теперь из (4.54) получаем формулы точечного преобразования

$$\begin{aligned} l &= p+2(2R_1-1)pq+2(2P_3-1)pr,\\ m &= q+2(2R_1-1)pq+2(2R_2-1)qr, \\ n &= r+2(2R_3-1)pr+2(2R_2-1)qr. \end{aligned} \tag{4.55}$$

С помощью громоздких выкладок можно показать, что *поверхность* $1-p-q-r=0$ *оказывается многочленом четвертой степени по* l, m *и* n.

(E) Шестигранник с криволинейными гранями

Ограниченная область в трехмерном пространстве, имеющая криволинейную границу, может быть разбита на конечное число шестигранных элементов с криволинейными гранями. Точечные преобразования, основанные на изопарамет-

[1]) При таком разбиении куски границы области дублируются примыкающими к границе гранями элементов наподобие того, как это происходит в апельсине.

рических аппроксимациях типа (4.35) и (4.36), можно использовать для перехода от произвольного шестигранника к единичному кубу в (p, q, r)-пространстве. В преобразованном пространстве вычисления тогда выполняются обычным образом. К сожалению, формулы преобразования, соответствующие (4.35) и (4.36), столь сложны, что из них невозможно получить хоть какие-то представления о форме криволинейной поверхности, порождаемой точечными преобразованиями. Единственным утешением служит то, что это можно сделать, если провести поверхность через большое число точек, лежащих на первоначальной криволинейной поверхности.

Замечание. Имеется большая потребность в базисных функциях, которые *точно* воспроизводят конкретные криволинейные стороны и поверхности в двумерном и трехмерном пространствах соответственно. Это особенно важно в задачах динамики невязкой жидкости, где «подделка» границы полностью изменяет поведение скорости вблизи границы. Хотя изопараметрический подход является улучшением по сравнению с заменой кривых линий и поверхностей прямыми линиями и плоскостями соответственно, он все же основан на точечных преобразованиях и поэтому дает только аппроксимацию исходных кривых и поверхностей. Предварительные результаты, основанные на геометрических рассмотрениях для нахождения точных базисных функций криволинейных элементов, можно найти у Уачспресса (1975), Маклеода (1977), Маклеода и Митчелла (1972) и Барнхилла и Грегори (1976a) и (1976b).

ГЛАВА 5

СХОДИМОСТЬ АППРОКСИМАЦИЙ

5.1. Введение

В гл. 3 было дано определение аппроксимации Галеркина. Было показано, что такая аппроксимация U удовлетворяет уравнению

$$a(U, V) = (f, V) \quad (\text{для всех } V \in K_N), \tag{5.1}$$

где K_N есть N-мерное подпространство пространства допустимых функций $\mathscr{H}$. Если предположить, что интегралы вычисляются точно, то анализ ошибок для кусочно-гладких аппроксимаций вида

$$U(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \varphi_i(\mathbf{x})$$

сводится к двум основным моментам:

(1) Доказательству того, что аппроксимация является наилучшей или почти наилучшей. В разд. 3.5 было показано, что аппроксимация Ритца будет наилучшей в энергетической норме, т. е.

$$\|u - U\|_A = \inf_{\tilde{u} \in K_N} \|u - \tilde{u}\|_A. \tag{5.2}$$

В общем же случае удается показать лишь то, что аппроксимация Галеркина является *почти наилучшей* в некоторой *соболевской норме*[1]): это означает, что

$$\|u - U\|_{r,R} \leqslant C \inf_{\tilde{u} \in K_N} \|u - \tilde{u}\|_{r,R} \tag{5.3}$$

при некоторых r, $C > 0$[2]).

(2) Оценке верхней границы правой части (5.3) в том специальном случае, когда элемент $\tilde{u} \in K_N$ интерполирует решение. Если K_N есть пространство конечноэлементных аппроксимаций, то обычно существуют целое $k = k(K_N)$ и постоянная $C = C(K_N)$, такие, что ошибка интерполяции огра-

[1]) Определение пространств и норм Соболева дается на стр. 114.

[2]) В этой главе через C обозначается положительная постоянная, своя для каждого конкретного случая.

ничена как

$$\| u - \tilde{u} \|_{r,\, R} \leqslant Ch^{k+1-r} \| u \|_{k+1,\, R} \quad (r = 0,\ 1,\ \ldots,\ k) \qquad (5.4)$$

при условии, что $u \in \mathscr{H}_2^{(k+1)}(R)$. Если предположить, что решение u удовлетворяет соотношению

$$a(u, v) = (f, v) \quad (\text{для всех } v \in \mathscr{H})$$

и является элементом соболевского пространства $\mathscr{H}_2^{(k+1)}(R)$, неравенства (5.3) и (5.4) могут быть объединены для получения оценки порядка сходимости аппроксимации Галеркина U при стремлении h к нулю.

Если, как это обычно имеет место, интегралы получаются численно по некоторой квадратурной формуле, то приближенное решение не определяется более соотношением (5.1), а определяется его приближенным аналогом

$$a_h(U_h, V) = (f, V)_h \quad (\text{для всех } V \in K_N). \qquad (5.5)$$

При этом оказывается возможным получить оценку порядка сходимости с помощью (5.3) и (5.4), если справедливо неравенство

$$\| U - U_h \|_{r,\, R} \leqslant Ch^s$$

при некотором $s > 0$. Аппроксимация границы также приводит к некоторому изменению уравнений системы (5.1), как и использование несогласованных элементов, т. е. недопустимых функций $U_h \notin \mathscr{H}$. Поэтому ясно, что анализ вопросов, порождаемых приближенным характером системы (5.5), имеет большое значение при *практическом* использовании метода конечных элементов.

Обозначения и вводные замечания

Билинейная форма a называется *эллиптичной* [1]) *в* $\mathscr{H}$, если существует такая постоянная $\gamma > 0$, что для всех $u \in \mathscr{H}$

$$a(u, u) \geqslant \gamma \| u \|^2.$$

По аналогии с гл. 1 назовем ее ограниченной, если существует такая постоянная $\alpha > 0$, что для всех $u, v \in \mathscr{H}$

$$| a(u, v) | \leqslant \alpha \| u \| \| v \|.$$

Большинство оценок погрешностей в этой главе выражается в терминах соболевских норм. Такая норма определяется как

$$\| u \|_{k,\, R}^2 = \sum_{|i| \leqslant k} \| D^i u \|_{\mathscr{L}_2(R)}^2,$$

[1]) Иногда говорят, что она коэрцитивна или положительно определена.

где в правой части использованы введенные ранее мультииндексные обозначения. Полезным оказывается также понятие полунормы, определяемой производными только одного порядка:

$$|u|^2_{k,R} = \sum_{|i|=k} \|D^i u\|^2_{\mathscr{L}_2(R)}.$$

Пространство Соболева $\mathscr{H}_2^{(k)}(R)$ состоит из всех тех функций, для которых соответствующая соболевская норма конечна. Через (u, v) будет обозначаться (если не оговорено ничего другого) скалярное произведение функций $u(\mathbf{x})$ и $v(\mathbf{x})$ в смысле пространства $\mathscr{L}_2(R)$:

$$(u, v) = \iint_R u(\mathbf{x})\, v(\mathbf{x})\, d\mathbf{x}.$$

Двойственное к $\mathscr{H}_2^{(k)}(R)$ пространство обозначается через $\mathscr{H}_2^{(-k)}(R)$, а соответствующая ему норма имеет вид

$$\|F\|_{-k,R} = \sup_{u \in \mathscr{H}_2^{(k)}(R)} \left\{ \frac{|F(u)|}{\|u\|_{k,R}} \right\}.$$

В разделе 5.4 (В), где элемент $F \in \mathscr{H}_2^{(-k)}(R)$ будет определен для некоторого $v \in \mathscr{H}_2^{(k)}(R)$ как

$$F(u) = (u, v) \quad (\text{при всех } u \in \mathscr{H}_2^{(k)}(R)),$$

можно положить, не внося двусмысленности в обозначения,

$$\|v\|_{-k,R} = \sup_{u \in \mathscr{H}_2^{(k)}(R)} \left\{ \frac{|(u, v)|}{\|u\|_{k,R}} \right\}.$$

Упражнение 1. Докажите, что для любого $u \in \mathscr{H}_2^{(k)}(R)$

$$\|u\|_{-k,R} \leqslant \|u\|_{\mathscr{L}_2(R)}.$$

В следующих неравенствах обозначаемые через C постоянные могут зависеть от области R и от пространства $\mathscr{H}$ допустимых функций, но не будут зависеть от самих рассматриваемых функций. Предполагается, что R является открытой ограниченной областью с гладкой (или кусочно-гладкой) границей ∂R, что $\bar{R} = R \cup \partial R$ и, если не оговорено противное, $R \subset \mathbb{R}^2$. Для достаточно гладких функций, принадлежащих, скажем, пространству $C^k(\bar{R})$, иногда будет использоваться *максимум-норма*

$$\|u\|_{(k)\bar{R}} = \max_{|i| \leqslant k} \{\|D^i u\|_{\mathscr{L}_\infty(\bar{R})}\}$$

вместе с соответствующей полунормой

$$|u|_{(k),\bar{R}} = \max_{|i|=k} \{\| D^i u \|_{\mathscr{L}_\infty(\bar{R})}\}.$$

Для любого $k > 0$ определим пространство $\mathring{\mathscr{H}}_2^{(k)}(R)$:

$$\mathring{\mathscr{H}}_2^{(k)}(R) = \left\{u: u \in \mathscr{H}_2^{(k)}(R),\ u = \frac{\partial u}{\partial n} = \ \ldots\ = \frac{\partial^{k-1} u}{\partial n^{k-1}} = 0 \text{ на } \partial R\right\}$$

и аналогично для любого конечномерного подпространства $K_N \subset \mathscr{H}_2^{(k)}(R)$ определим

$$\mathring{K}_N = K_N \cap \mathring{\mathscr{H}}_2^{(k)}(R);$$

через $\bar{K}_N$ обозначается дополнение к $\mathring{K}_N$ в K_N. Например, для любого подпространства K_N конечных элементов $\bar{K}_N$ будет подпространством, образованным только теми базисными функциями, которые соответствуют граничным узлам, т. е. обращаются в нуль во всех внутренних узлах.

Для краткости изложения различные дополнительные ограничения на область R при формулировке следующих лемм будут опущены. Эти ограничения не являются особенно обременительными и обычно выполнены, если граница будет достаточно гладкой между угловыми точками. Интересующийся читатель может обратиться к литературе, на которую в каждом случае имеются детальные ссылки. Предполагается, в частности, что лемма Соболева (см., например, Агмон, 1965, стр. 32 или Иосида, 1965, стр. 242) применима к области $R \subset \mathbb{R}^m$, т. е. что при $k > r + m/2$

$$\| u \|_{(r),(\bar{R})} \leqslant C \| u \|_{k,R}.$$

Лемма 5.1. *При всех* $u \in \mathring{\mathscr{H}}_2^{(1)}(R)$

$$\| u \|_{\mathscr{L}_2(R)} \leqslant C |u|_{1,R}.$$

Если λ (>0) есть минимальное собственное значение для задачи, определяемой дифференциальным оператором Лапласа и однородными граничными условиями Дирихле, то $C = 1/\lambda$ (Курант и Гильберт, 1953, стр. 386—392). Из леммы 5.1 следует, что для $\mathring{\mathscr{H}}_2^{(1)}(R)$ величины $\|.\|_{1,R}$ и $|.|_{1,R}$ будут эквивалентными нормами. Лемма 5.1 является частным случаем более общих результатов:

Лемма 5.2 (Обэн, 1972, стр. 173).

$$\| u \|_{\mathscr{L}_2(R)} \leqslant C |u|_{k,R} \quad (\textit{для всех } u \in \mathring{\mathscr{H}}_2^{(k)}(R)).$$

Лемма 5.3 (Нечас, 1967, стр. 18).

$$\|u\|^2_{k,R} \leqslant C\left\{|u|^2_{k,R} + \sum_{|i|<k}\left|\int\int D^i u\, d\mathbf{x}\right|^2\right\} \quad (\text{для всех } u \in \mathscr{H}_2^{(k)}(R)).$$

Теорема 5.1 (лемма Брамбла — Гильберта; Брамбл и Гильберт, 1970; Сьярле и Равьяр, 1972а). *Пусть элемент* $F \in \mathscr{H}_2^{(-k-1)}(R)$ *таков, что* $F(p) = 0$ *для всех* $p \in P_k$. *Тогда существует такая постоянная* $C = C(R)$, *что*

$$|F(u)| \leqslant C\|F\|_{-k-1,R}|u|_{k+1,R} \quad (\textit{для любого } u \in \mathscr{H}_2^{(k+1)}(R)).$$

Доказательство. Можно показать (упражнение 2), что для любого $u \in \mathscr{H}_2^{(k+1)}(R)$ существует такой полином $p = p(u) \in P_k$, что

$$\int\int_R D^i(u+p)\, d\mathbf{x} = 0 \quad (|i| \leqslant k).$$

Применяя теперь лемму 5.3, получим

$$\|u+p\|^2_{k+1,R} \leqslant C|u+p|^2_{k+1,R} = C|u|^2_{k+1,R}.$$

Тогда в силу линейности функционала F

$$F(u) = F(u+p)$$

и после объединения всех этих результатов будем иметь

$$|F(u)| \leqslant \|F\|_{-k-1,R}\|u+p\|_{k+1,R} \leqslant C\|F\|_{-k-1,R}|u|_{k+1,R}.$$

Упражнение 2. Докажите индукцией по k, что для любого $u \in \mathscr{H}_2^{(k+1)}(R)$ существует такой полином $p = p(u) \in P_k$, что

$$\int\int_R D^i(u+p)\, d\mathbf{x} = 0\,(|i| \leqslant k).$$

Чаще всего лемма Брамбла — Гильберта используется для получения оценок для билинейных форм. Например, пусть $\mathscr{H}$ есть гильбертово пространство и F — ограниченная билинейная форма с аргументами из $\mathscr{H}_2^{(k+1)}(R)$ и $\mathscr{H}$, т. е. $F \in \mathscr{L}(\mathscr{H}_2^{(k+1)}(R) \times \mathscr{H}; \mathbb{R})$. Тогда если

$$F(u, v) = 0$$

для всех $v \in \mathscr{H}$ при $u \in P_k$ и функционал $F_1 \in \mathscr{H}_2^{(-k-1)}(R)$ определен для некоторого $v \in \mathscr{H}$ в виде

$$F_1(u) = F(u, v) \quad (\text{для всех } u \in \mathscr{H}_2^{(k+1)}(R)),$$

то лемма Брамбла — Гильберта приводит к оценке

$$|F_1(u)| \leqslant C \|F_1\|_{-k-1,R} |u|_{k+1,R}.$$

В разд. 1.2 (упражнение 13) показано, что для такого функционала

$$\|F_1\| \leqslant \|F\| \|v\|_{\mathcal{H}},$$

и объединение двух этих результатов приводит к неравенству

$$|F(u, v)| \leqslant C \|F\| \|v\|_{\mathcal{H}} |u|_{k+1,R}. \tag{5.6}$$

Упражнение 3. Используя лемму 5.3, доказать, что если $F \in \mathcal{L}\left(\mathcal{H}_2^{(k+1)}(R) \times \mathcal{H}_2^{(r+1)}(R); \mathbb{R}\right)$ такой, что

$$F(u, v) = 0 \quad \begin{cases} \text{для всех } u \in \mathcal{H}_2^{(k+1)}(R), \text{ если } v \in P_r \\ \text{для всех } v \in \mathcal{H}_2^{(r+1)}(R), \text{ если } u \in P_k, \end{cases}$$

то

$$|F(u, v)| \leqslant C \|F\| |u|_{k+1,R} |v|_{r+1,R}.$$

Регулярные преобразования

Обычно базисные функции определяются на стандартном элементе T_0, который может быть единичным квадратом или прямоугольным треугольником с единичными катетами, а затем вводится точечное преобразование для построения базисных функций на произвольном элементе T (ср. с гл. 4). Поэтому естественно получать оценки погрешностей на стандартном элементе, если только они допускают обобщение на произвольные элементы.

В случае криволинейных элементов такое точечное преобразование обычно разбивается на два этапа путем введения промежуточного элемента T'. Этот промежуточный элемент имеет те же вершины, что и криволинейный элемент T, но стороны его уже прямолинейны. Поэтому T_0 отображается на T' линейным преобразованием $\mathbf{l} = \mathbf{F}_0(\mathbf{p})$, таким, что

$$t = t_3 + (t_1 - t_3)\, p + (t_2 - t_3)\, q \quad (t = l, m),$$

где $\mathbf{p} = (p, q) \in T_0$ и $\mathbf{l} = (l, m) \in T'$. Тогда отображение T' на криволинейный элемент T можно рассматривать как нелинейное возмущение линейного преобразования. Если $\mathbf{x} = (x, y) \in T$, то полное преобразование запишется в виде

$$\mathbf{x} = \mathbf{F}(\mathbf{p}) = \mathbf{F}_0(\mathbf{p}) + \mathbf{F}_1(\mathbf{p}),$$

где $\mathbf{F}_0$ есть линейное преобразование, а $\mathbf{F}_1$ — нелинейный поправочный член. Все рассмотренные в гл. 4 криволинейные

элементы могут быть представлены таким образом. Заметим, что *построение* некоторых элементов в разд. 4.6 проводилось с помощью другой последовательности преобразований, в которой промежуточный элемент T' имел криволинейные стороны и те же вершины, что и стандартный треугольник T_0.

Чтобы неравенства вида (5.6) можно было использовать для получения оценок порядка сходимости конкретных конечноэлементных аппроксимаций, часто бывает необходимо отобразить $\mathscr{H}_2^{(r)}(T_0)$ на $\mathscr{H}_2^{(r)}(T)$ и обратить это отображение. Поэтому мы будем считать отображение $\mathbf{F}$ настолько гладким, чтобы из $v \in \mathscr{H}_2^{(r)}(T)$, следовало бы, что $v \circ \mathbf{F} \in \mathscr{H}_2^{(r)}(T_0)$, где *сложный* (или *составной*) оператор $v \circ \mathbf{F}$ определен как

$$(v \circ \mathbf{F})(\mathbf{p}) = v(\mathbf{F}(\mathbf{p})) \quad \text{(для всех } \mathbf{p} \in T_0\text{)}.$$

Для упрощения обозначений мы часто будем писать v вместо $v \circ \mathbf{F}$ и в случае необходимости $v(p)$ или, например, $v(\mathbf{x})$, чтобы исключить возможную двусмысленность. Предположим также, что обратное преобразование $\mathbf{F}^{-1}$ настолько гладко, что при $v(\mathbf{p}) \in \mathscr{H}_2^{(r)}(T_0)$ выполняется включение $v(\mathbf{x}) \in \mathscr{H}_2^{(r)}(T)$.

Условие 5.1 (условие регулярности). *Если* $v \in \mathscr{H}_2^{(r)}(T)$ *и диаметр* [1]) *элемента* T *есть* h, *то существуют такие постоянные* C_0, C_1 и C_2, что

$$|v|_{r, T_0} \leqslant C_1 \{\inf_{\mathbf{p} \in T_0} J(\mathbf{p})\}^{-1/2} h^r \|v\|_{r, T}, \tag{5.7a}$$

$$|v|_{r, T_0} \geqslant C_2 \{\sup_{\mathbf{p} \in T_0} J(\mathbf{p})\}^{-1/2} h^r |v|_{r, T} \tag{5.7b}$$

и

$$0 < \frac{1}{C_0} \leqslant \left\{ \frac{\sup J(\mathbf{p})}{\inf J(\mathbf{p})} \right\} \leqslant C_0, \tag{5.7c}$$

где J *обозначает якобиан преобразования* $\mathbf{x} = \mathbf{F}(\mathbf{p})$.

Заметим, что это условие предполагает положительность якобиана для всех $\mathbf{p} \in T_0$. В гл. 4 было показано, что якобиан всегда положителен, если отсутствуют так называемые запрещенные элементы.

---

[1]) Диаметром треугольного элемента является длина его наибольшей стороны, диаметром четырехугольного элемента является длина его наибольшей диагонали.

Если все сводится к линейному преобразованию $\mathbf{x} = \mathbf{F}_0(\mathbf{p})$, то якобиан будет константой и

$$J_0 = \det\begin{bmatrix} \frac{\partial x}{\partial p} & \frac{\partial x}{\partial q} \\ \frac{\partial y}{\partial p} & \frac{\partial y}{\partial q} \end{bmatrix} = \det\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix},$$

и поэтому оценка (5.7c) становится тривиальной.

Условие регулярности выполняется для определенных нелинейных преобразований вида

$$\mathbf{x} = \mathbf{F}_0(\mathbf{p}) + F_1(\mathbf{p}),$$

где $\mathbf{F}_1$ есть «малое» возмущение (см. упражнения 9 и 10). Если возмущающий член удовлетворяет некоторым условиям, то можно показать, что якобиан имеет вид

$$J(\mathbf{p}) = J_0 + J_1(\mathbf{p}),$$

где величина J_1 также «мала», и поэтому выполняется условие 5.1 (Сьярле и Равьяр, 1972; Зламал, 1974).

Упражнение 4. Докажите, что из (5.7b) следует, что если $h < 1$, то

$$\| v \|_{r, T_0} \geqslant C \| v \|_{r, T} h^r \{ \sup_{\mathbf{p} \in T} J(\mathbf{p}) \}^{-1/2}.$$

Упражнение 5. Докажите, что для любого $v \in \mathscr{L}_2(T)$

$$\| Jv \|_{\mathscr{L}_2(T_0)} \leqslant \{ \sup_{\mathbf{p} \in T} J(\mathbf{p}) \}^{1/2} \| v \|_{\mathscr{L}_2(T)}.$$

Сочетание леммы Брамбла — Гильберта с условием регулярности применяется главным образом (но не исключительно) для оценки погрешности интерполяции. Столь же успешно это сочетание можно использовать для оценки ошибок, возникающих в результате применения численных квадратур, основанных на конечноэлементных разбиениях области интегрирования.

Упражнение 6. (I) Пусть $E_0(v)$ есть ошибка численного интегрирования на стандартном элементе для функции $v \in \mathscr{H}_2^{(k+1)}(T_0)$, и пусть квадратурная формула точна для полиномов степени не выше k. Покажите с помощью леммы Брамбла — Гильберта, что

$$| E_0(v) | \leqslant C | v |_{k+1, T_0}.$$

(II) Пусть теперь такая формула преобразована для интегрирования на произвольном элементе, полученном линейным преобразованием из стандартного элемента. Используя

условие регулярности, покажите, что ошибка интегрирования для $v \in \mathscr{H}_2^{(k+1)}(T)$ может быть представлена в виде $E_0(J_0 v)$ и что

$$|E_0(J_0 v)| \leqslant Ch^{k+1} |v|_{k+1, T}.$$

Упражнение 7. (I) Пусть $u \in \mathscr{H}_2^{(k+1)}(T_0)$, $w \in P_r$ и $E_0(uw)$ обозначает ошибку численного интегрирования на стандартном элементе произведения uw. Введем еще билинейную форму $E_1 \in \mathscr{L}(\mathscr{H}_2^{(k+1)}(T_0) \times \mathscr{L}_2(T_0); \mathbb{R})$ как

$$E_1(u, w) = E_0(uw)$$

и предположим, что квадратурная формула точна для полиномов степени не выше $r + k$. Докажите, что тогда найдется такая постоянная C, что

$$|E_0(uw)| \leqslant C \|w\|_{\mathscr{L}_2(T_0)} |u|_{k+1, T_0}.$$

(*Указание.* Доказательство аналогично выводу неравенства (5.6).)

(II) Пусть теперь такая квадратурная формула преобразована для интегрирования на произвольном элементе с помощью линейного преобразования. Покажите, что тогда ошибка численного интегрирования произведения двух функций $u \in \mathscr{H}_2^{(k+1)}(T)$ и $w \in P_r$ может быть представлена в виде $E_0(J_0 uw)$ и что

$$|E_0(J_0 uw)| \leqslant Ch^{k+1} |u|_{k+1, T} \|w\|_{\mathscr{L}_2(T)}.$$

В упражнениях 6 и 7 предполагалась линейность преобразований, и поэтому $J_0 w(\mathbf{p}) \in P_r$, когда $w(\mathbf{x}) \in P_r$. При использовании нелинейных преобразований, как это будет в разделе 5.4(A), необходимо рассматривать такие функции $w(\mathbf{x})$, для которых $w(\mathbf{p}) J(\mathbf{p})$ является полиномом. Дальнейшие детали по поводу видоизменения квадратурных формул для интегрирования на произвольных элементах изложены в разделе 5.4(A).

Упражнение 8. Докажите, что если $\mathbf{x} = \mathbf{F}_0(\mathbf{p})$, то найдутся такие постоянные C_1 и C_2, что

$$\iint_{T_0} \left(\frac{\partial v}{\partial t}\right)^2 d\mathbf{p} \leqslant C_1 h^2 \iint_T \left\{\left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right\} J_0 \, d\mathbf{x} \quad (t = p, q)$$

и

$$\iint_{T_0} \left(\frac{\partial v}{\partial t}\right)^2 d\mathbf{p} \geqslant C_2 h^2 \iint_T \left\{\left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right\} J_0 \, d\mathbf{x} \quad (t = p, q).$$

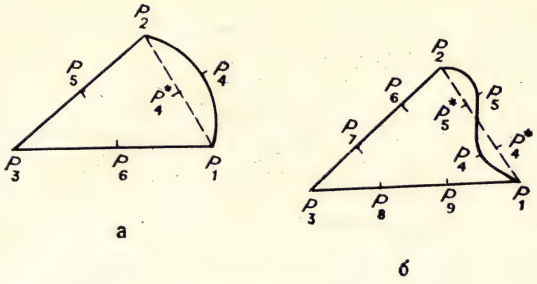


Рис. 24.

Докажите далее по индукции, что неравенства (5.7a) и (5.7b) останутся верными при $r = 1, 2, \ldots$, если в них заменить $\|v\|_{r,\tau}$ на $|v|_{r,\tau}$.

Упражнение 9. (a) Покажите, что если точки P_5 и P_6 являются серединами сторон P_2P_3 и P_1P_3 соответственно, то квадратичные изопараметрические элементы с одной криволинейной стороной (см. рис. 24(a)) задаются преобразованием

$$t = t_3 + (t_1 - t_3)\,p + (t_2 - t_3)\,q + \left(t_4 - \frac{t_1 + t_2}{2}\right) 4pq \quad (t = x,\ y)$$

(ср. с разд. 4.6).

(b) Покажите, что для такого преобразования

$$J(\mathbf{p}) = J_0 + J_1(\mathbf{p}),$$

где

$$J_1(\mathbf{p}) = 4\left(x_4 - \frac{x_1 + x_2}{2}\right)\{(y_2 - y_3)\,q - (y_1 - y_3)\,p\} + \\ + 4\left(y_4 - \frac{y_1 + y_2}{2}\right)\{(x_1 - x_3)\,p - (x_2 - x_3)\,q\}$$

Найдите достаточные условия, при которых

$$0 < J_0 - Ch^3 \leqslant J \leqslant J_0 + Ch^3,$$

т. е. $J = 0(h^2)$. Сравните ваши результаты с (5.19).

Зламал (1973) предложил для треугольных элементов с одной криволинейной стороной другой подход, который приводит к аналогичным оценкам.

Упражнение 10. Покажите, что если точки P_6 и P_7, P_8 и P_9 делят на три равные части стороны P_2P_3 и P_1P_3 соответственно (см. рис. 24(b)), то преобразование, задающее кубические изопараметрические элементы, содержит нелиней-

ные члены

$$27pq(1-p-q)\left(t_{10}-\frac{t_1+t_2+t_3}{3}\right)+$$

$$+\frac{9}{2}pq\left[(3p-1)\left(t_4-\frac{2t_1+t_2}{3}\right)+(3q-1)\left(t_5-\frac{2t_2+t_1}{3}\right)\right]$$

$$(t=x,\ y).$$

Полные аппроксимации

Условие регулярности содержит предположения о свойствах преобразования, переводящего произвольный элемент в стандартный элемент. При объединении этих предположений с леммой Брамбла — Гильберта можно получить оценки типа (5.4), т. е. определить порядок сходимости конечноэлементной аппроксимации. Поэтому основное значение леммы Брамбла — Гильберта состоит в получении оценок для ошибок интерполяции. Лемма одинаково хорошо может использоваться при оценке ошибок для любой формы аппроксимации, представимой как проекция на пространство кусочных полиномов. Чтобы применить лемму, сначала необходимо сделать некоторые предположения относительно типов функций, лежащих в основе конечноэлементной аппроксимации.

Для любого элемента T обозначим через $K_{[T]}$ пространство, определяемое теми пробными функциями, которые отличны от нуля на T. Другими словами, $K_{[T]}$ есть сужение пространства пробных функций относительно элемента T. Обозначим через $K_{[0]}$ пространство таких функций $v(\mathbf{p})$, для которых $v(\mathbf{x})\in K_{[T]}$. Это пространство $K_{[0]}$ имеет особо важное значение при анализе методов конечных элементов. Порядок метода определяется максимальной степенью полинома (по $\mathbf{p}$), для которого ошибка аппроксимации функцией из $K_{[0]}$ равна нулю. В общем случае эта степень совпадает с таким максимальным k, для которого $P_k\subset K_{[0]}$.

Например, если рассматривается аппроксимация кусочными кубическими полиномами (Лагранжа или Эрмита) на треугольной сетке (разд. 4.1), то преобразование F будет линейным и $K_{[0]}$, $K_{[T]}=P_3$. Для аппроксимации (4.14) (определяемые 18 параметрами полиномы пятой степени со сшивкой в C^1) преобразование также будет линейным, но $P_4\subset K_{[0]}$, $K_{[T]}\subset P_5$ (строгое вложение), а для биквадратичной изопараметрической аппроксимации (4.29) $P_2\subset K_{[0]}\subset P_4$ (снова строгое вложение), но преобразование уже не будет линейным, т. е. $K_{[T]}$ не будет полиномиальным подпространством.

Для каждого элемента T введем проекцию Π_T на $K_{[T]}$, так что для любой достаточно гладкой функции u функция

$\Pi_T u(\mathbf{x})$ будет интерполировать $u(\mathbf{x})$ на T. Интерполяция в этом контексте означает согласование всех узловых параметров, которыми определяется конечноэлементная аппроксимация. Для функций, заданных на стандартном элементе, можно определить отображение Π на $K_{[0]}$ как

$$\Pi(v \circ \mathbf{F}) = (\Pi_T v) \circ \mathbf{F}.$$

Условие 5.2 (условие полноты). *Для любого элемента T диаметра h существует такое $k > 0$, что $P_k \subset K_{[0]}$ и для проекции Π нормы $\|I - \Pi\|$ равномерно ограничены при всех h.*

Пример того, что нормы $\|I - \Pi\|$ не будут равномерно ограничены, доставляют треугольные элементы, когда нормальная производная в точке стороны является параметром, а значение функции и тангенциальная производная — нет, и некоторые треугольники стремятся к вырожденным при измельчении сетки, т. е. есть такие элементы, у которых наименьший угол стремится к нулю (Брамбл и Зламал, 1970). Мы снова вернемся к этому примеру в конце разд. 5.3.

В параграфе 5.3 предполагается, что можно определить пространство $\mathscr{H}_2^{(k)}(R)$ для нецелых k так же, как и для целых. Детальное обсуждение свойств такого пространства и смысла теоремы о следе выходит за рамки этой книги, и мы отсылаем интересующегося читателя к работам Обэна (1972), Лионса и Мадженеса (1972) или Нечаса (1967). Краткие сведения о наиболее важных свойствах этих пространств и об их приложениях можно найти в первой главе книги Варги (1971).

5.2. Сходимость аппроксимаций Галеркина

Продолжением результатов разд. 3.5 является доказательство того, что аппроксимация Галеркина для линейных задач в общем случае является *почти наилучшей* в смысле того определения, которое было дано в предыдущем параграфе. Основой такого доказательства служит обобщение леммы Лакса — Мильграма (Иосида, 1965, стр. 134):

Теорема 5.2 (Обэн, 1972, стр. 40). *Если $\mathscr{H}$ есть гильбертово пространство, $l \in \mathscr{H}'$ и a — эллиптичная в $\mathscr{H}$ ограниченная билинейная форма, то существует единственный вектор $u \in \mathscr{H}$, такой, что*

$$a(u, v) = l(v) \quad (\textit{для всех } v \in \mathscr{H}).$$

Для данного N-мерного подпространства K_N существует единственный вектор $U \in K_N$, такой, что

$$a(U, V) = l(V) \quad (\textit{для всех } V \in K_N).$$

Более того,

$$\|u - U\|_{\mathscr{H}} \leqslant C \inf_{\tilde{u} \in K_N} \|u - \tilde{u}\|_{\mathscr{H}}.$$

Аналогичный результат может быть получен для некоторых нелинейных задач, допускающих применение теории монотонных операторов (Варга, 1971, гл. 4).

В качестве примера результатов, которые могут быть получены с помощью теоремы 5.2, рассмотрим аппроксимацию Ритца решения уравнения

$$\frac{\partial}{\partial x}\left(d_1(x, y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(d_2(x, y)\frac{\partial u}{\partial y}\right) + f(x, y) = 0 \quad ((x, y) \in R) \tag{5.8}$$

с граничным условием

$$u(x, y) = 0 \quad ((x, y) \in \partial R), \tag{5.9}$$

где предполагается существование таких постоянных δ и Δ, что

$$0 < \delta \leqslant d_1(x, y),\ d_2(x, y) \leqslant \Delta \quad ((x, y) \in R).$$

Для этого примера

$$l(v) = (f, v)$$

и

$$a(u, v) = \iint_R \left\{ d_1 \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + d_2 \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right\} dx\, dy.$$

Так как

$$|a(u, v)| \leqslant \Delta \iint_R \left\{ \left|\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\right| + \left|\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right| \right\} dx\, dy,$$

билинейная форма будет ограниченной, а поскольку

$$a(u, u) \geqslant \delta |u|_{1, R},$$

то из леммы 5.1 следует, что она будет эллиптичной в $\mathring{\mathscr{H}}_2^{(1)}(R)$.

Следствие теоремы 5.2. *Если и есть решение задачи* (5.8)—(5.9), *то аппроксимация Ритца* $U \in \mathring{K}_N \subset \mathring{\mathscr{H}}_2^{(1)}(R)$ *такова, что*

$$\|u - U\|_{1, R} \leqslant C \inf_{\tilde{u} \in \mathring{K}_N} \|u - \tilde{u}\|_{1, R}.$$

Упражнение 11. Докажите, что если (I) u есть решение уравнения (5.8) с граничным условием $u = g$ на ∂R и (II) $w \in \mathscr{H}_2^{(1)}(R)$ — *любая* такая функция, что $w = g$ на ∂R, то аппроксимация Ритца $U + w$ при $U \in \mathring{K}_N \subset \mathring{\mathscr{H}}_2^{(1)}(R)$ такова, что

$$\|u - (U + w)\|_{1, R} \leqslant C \inf_{\tilde{u} \in \mathring{K}_N} \|u - (\tilde{u} + w)\|_{1, R}$$

Упражнение 12. Докажите, что если u есть решение уравнения (5.8) с граничным условием $\frac{\partial u}{\partial n} = g$ на ∂R, то существует такая аппроксимация Ритца $U \in K_N \subset \mathscr{H}_2^{(1)}(R)$, что

$$\|u - U\|_{1,R} \leqslant C \inf_{\tilde{u} \in K_N} \|u - \tilde{u}\|_{1,R}.$$

Отметим, что для существования решения необходимо предположение о совместимости f и g. Например, если в (5.8) $d_1 = d_2 = 1$, то мы предположим, что

$$\iint_R f\,dx\,dy = -\int_{\partial R} g\,d\sigma,$$

и так как решение определяется однозначно с точностью до постоянного слагаемого, то это последнее можно выбрать так, чтобы

$$\iint_R u\,dx\,dy = 0$$

(Нечас, 1967, стр. 256). Следовательно, для получения результата можно применить лемму 5.3 при $k = 1$.

Если необходимо аппроксимировать граничные условия базисными функциями, принимающими на границе ненулевые значения, то указать границу ошибки еще возможно. Если

$$U = U_0 + \bar{U},$$

где $U_0 \in \mathring{K}_N \subset \mathring{\mathscr{H}}_2^{(1)}(R)$, то $\bar{U}$ содержит только такие базисные функции, которые могут принимать ненулевые значения на границе, и полностью определяется граничными данными. Из определений параграфа 5.1 следует, что $\bar{U} \in \bar{K}_N$.

Лемма 5.4. *Если u есть решение уравнения* (5.8) *с граничным условием $u = g$ на ∂R и $\bar{U} \in \bar{K}_N$ выбрано так, что $\bar{U}(x, y)$ $((x, y) \in \partial R)$ есть фиксированная аппроксимация функции g, то конечноэлементная аппроксимация $U = U_0 + \bar{U}$ при $U_0 \in \mathring{K}_N \subset \mathring{\mathscr{H}}_2^{(1)}(R)$ такова, что*

$$|u - U|_{1,R} \leqslant C\,|u - (\tilde{u} + \bar{U})|_{1,R} (\textit{для всех } \tilde{u} \in \mathring{K}_N).$$

Доказательство. Так как

$$u - (\tilde{u} + \bar{U}) = u - U + U_0 - \tilde{u}$$

и $U_0 - \tilde{u} \in \mathring{K}_N$, то результат непосредственно следует из леммы 5.3 и определений u и U.

Если мы предположим, что существует гладкое продолжение g в R, то лемму 5.4 можно использовать также для получения оценок в терминах пространства $\mathscr{H}_2^{(1)}(R)$. Существует по крайней мере одно такое продолжение — именно само решение u.

Теорема 5.3 (Файервезер, 1972, стр. 45). *Пусть u есть решение уравнения* (5.8) *с граничным условием $u = g$ на ∂R и $w \in \mathscr{H}_2^{(1)}(R)$ — любое гладкое продолжение g в R. Тогда конечноэлементная аппроксимация вида $U = U_1 + W$, где $W \in K_N$ есть аппроксимация w, а $U_1 \in \overset{\circ}{K}_N$, будет такой, что*

$$\| u - (U_1 + W) \|_{1, R} \leqslant C \{ \| u - (\tilde{u} + W) \|_{1, R} + \| w - W \|_{1, R} \} \quad (5.10)$$

для любого $\tilde{u} \in \overset{\circ}{K}_N$.

Отметим, что если, например, граничное условие получается с помощью интерполяции, то правая часть (5.10) состоит из ошибки аппроксимации функции $u - w \in \overset{\circ}{\mathscr{H}}_2^{(1)}(R)$ и ошибки интерполяции функции w, которая отлична от нуля на границе.

Аппроксимация границы и численное интегрирование

Построение конечноэлементной аппроксимации для задач с интерполированными граничными условиями — как это только что было — это одно из основных *нарушений вариационных принципов* (Стренг, 1972), на которые приходится постоянно идти при решении практических задач. Другие нарушения таковы: (I) искажение положения границы; (II) использование численного интегрирования для вычисления скалярных произведений и (III) применение несогласованных элементов. Несогласованные элементы будут детально рассмотрены в разд. 7.2. Если применяется любой из этих приемов, то приближенное решение не лежит более в K_N и не удовлетворяет условию

$$a(U, V) = (f, V) \quad (\text{для всех } V \in K_N). \quad (5.11)$$

Вместо этого решение $U_h \in K_h$ и удовлетворяет условию

$$a_h(U_h, V_h) = (f, V_h)_h \quad (\text{для всех } V_h \in K_h), \quad (5.12)$$

где обе части (5.12) и вид нового M-мерного пространства K_h, которое может содержать функции, не являющиеся допустимыми для изложенных в гл. 3 классических вариационных методов, зависят от характера нарушений вариационных принципов. Обычно бывает, что $M \geqslant N$ и $K_h \supseteq K_N$. Если

$K_h = K_N$, как в том случае, когда единственным отличием от классической формы вариационного метода является применение численного интегрирования, то решения обеих систем (5.11) и (5.12) будут линейными комбинациями функций $\varphi_i(x)$ $(i = 1, \ldots, N)$. Коэффициенты такой комбинации для задачи (5.11) определяются из системы

$$G\alpha = \mathbf{b}, \tag{5.13}$$

где *матрица жесткости* $G = \{a(\varphi_i, \varphi_j)\}$ и $\mathbf{b} = \{(f, \varphi_i)\}$ $(i, j = 1, \ldots, N)$. Для приближенной задачи (5.12)

$$G_h \alpha_h = \mathbf{b}_h, \tag{5.14}$$

где $G_h = \{a_h(\varphi_i, \varphi_j)\}$ и $\mathbf{b}_h = \{(f, \varphi_i)_h\}$. Можно сравнить конкретные элементы G_h и $\mathbf{b}_h$ с соответствующими элементами G и $\mathbf{b}$ и получить оценку нормы $\|U - U_h\|$ известным из линейной алгебры стандартным методом малых возмущений. К сожалению, этот метод в большинстве случаев позволяет получить лишь грубую оценку сверху для нормы $\|U - U_h\|$ (Фикс, 1972). Если предположить, что приближенная билинейная форма эллиптична в $\mathscr{H}$ и ограничена, то тогда теорема 5.2 не только обеспечивает существование единственного решения, но и позволяет оценить степень приближения.

Упражнение 13. Покажите, что если a_h есть эллиптичная в $\mathscr{H}$ билинейная форма, то

$$a_h(U - U_h, W_h) = (a_h - a)(U, W_h) + (f, W_h) - (f, W_h)_h$$

для любого $W_h \in K_h$. Тем самым докажите, что

$$\|U - U_h\| \leqslant C \sup_{W_h \in K_N} \left\{ \frac{|(a_h - a)(U, W_h)| + |(f, W_h) - (f, W_h)_h|}{\|W_h\|} \right\}, \tag{5.15}$$

и получите аналогичную оценку для $\|u - U_h\|_{\mathscr{H}}$.

Упражнение 14. Покажите, что если a_h есть ограниченная билинейная форма, то

$$|a_h(U_h - V_h, W_h)| \leqslant \alpha \|u - V_h\| \|W_h\| + |(f, W_h)_h - a_h(u, W_h)|$$

для любых $V_h, W_h \in K_h$. Докажите далее, что если a_h к тому же эллиптична в $\mathscr{H}$, то

$$\|U_h - V_h\| \leqslant \frac{\alpha}{\gamma} \|u - V_h\| + \sup_{W_h \in K_h} \left\{ \frac{1}{\gamma} \frac{|(f, W_h)_h - a_h(u, W_h)|}{\|W_h\|} \right\}$$

для любого $V_h \in K_h$, и существует такое $C > 0$, что

$$\|u - U_h\| \leqslant C \left\{ \inf_{V_h \in K_h} \|u - V_h\| + \sup_{W_h \in K_h} \left\{ \frac{|(f, W_h)_h - a_h(u, W_h)|}{\|W_h\|} \right\} \right\}. \tag{5.16}$$

Отметим, что оценкой (5.15) можно пользоваться только тогда, когда U_h в некотором смысле близко к классической аппроксимации Галеркина. Это будет так, если $K_h = K_N$ или $K_h = K_N \oplus \{\varphi_{N+1}, \ldots, \varphi_M\}$, где дополнительные $\varphi_i(\mathbf{x})$ $(i = N+1, \ldots, M)$ обладают такими специальными свойствами, которые исключают их из числа допустимых для классической аппроксимации функций: например, они могут быть несогласованными. Оценка (5.16), наоборот, применима тогда, когда наше приближение существенно отличается от любой классической аппроксимации, как, например, в случае несогласованности всех элементов (разд. 5.4 (E)).

Аналогичным образом метод малых возмущений может быть применен и тогда, когда (5.5) является *системой вариационных разностных уравнений* (см., например, работу Демьяновича, 1964).

Упражнение 15. Докажите, что если a_h и a такие билинейные формы, что a_h эллиптична в K_h, а a ограничена, то

$$\|u - U_h\| \leqslant C \left\{ \inf_{V_h \in K_h} \left[\|u - V_h\| + \sup_{W_h \in K_h} \left\{ \frac{|(a_h - a)(V_h, W_h)|}{\|W_h\|} \right\} \right] + \right.$$
$$\left. + \sup_{W_h \in K_h} \left\{ \frac{|(l, W_h)_h - (f, W_h)|}{\|W_h\|} \right\} \right\} \qquad (5.17)$$

5.3. Ошибки аппроксимации

Как показано в предыдущем разделе, аппроксимации Галеркина всегда *почти оптимальны* в некоторой норме. В частности, приближенные решения для задач второго порядка таковы, что

$$\|u - U\|_{1,R} \leqslant C \inf_{\tilde{u} \in K_N} \|u - \tilde{u}\|_{1,R}.$$

Оценки ошибки также могут быть получены в терминах других норм, в частности норм пространств $\mathscr{L}_2(R)$ и $\mathscr{L}_\infty(\bar{R})$. Оценки в $\mathscr{L}_\infty(\bar{R})$ позволяют также доказать так называемое свойство *сверхсходимости* некоторых методов Галеркина, т. е. наличие в узлах более высокого порядка аппроксимации по сравнению с неузловыми точками (Дуглас, Дюпон и Уилер, 1974). Более подробно с этим вопросом можно познакомиться в работе де Бура (1974).

Пусть T_0 есть стандартный элемент. Тогда большинство результатов о сходимости может быть получено с помощью следующей леммы о полиномиальной аппроксимации на T_0:

Лемма 5.5. *Пусть* $\Pi \in \mathscr{L}\left(\mathscr{H}_2^{(k+1)}(T_0);\ \mathscr{H}_2^{(r)}(T_0)\right)$ $(k \geqslant r)$ *есть проекция на* $K_{[0]}$, *где* $\mathscr{H}_2^{(r)}(T_0) \supset K_{[0]} \supset P_k$. *Тогда*

$$\| v - \Pi v \|_{r,\,T} \leqslant C \| I - \Pi \| \, | v |_{k+1,\,T_0}$$

для всех $v \in \mathscr{H}_2^{(k+1)}(T_0)$.

Это значит, что если интерполяция точна для полиномов степени не выше k, то ошибка интерполяции может быть выражена через $(k+1)$-е производные интерполируемой функции. Операторы Π и $I - \Pi$ переводят элементы пространства $\mathscr{H}_2^{(k+1)}(T_0)$ в элементы пространства $\mathscr{H}_2^{(r)}(T_0)$, поскольку при конкретном применении леммы r зависит от порядка дифференциального уравнения, а k зависит от свойств пробных функций. Значение $\|I - \Pi\|$ зависит от значений r и k, но предполагается, что эти нормы равномерно ограничены и, вообще говоря, необязательно знать точное значение нормы.

Доказательство леммы 5.5. Для любого $G \in \mathscr{H}_2^{(-r)}(T_0)$ определим $F \in \mathscr{H}_2^{(-k-1)}(T_0)$ так, что при любом $v \in \mathscr{H}_2^{(k+1)}(T_0)$

$$F(v) = G([I - \Pi] v).$$

Тогда, применяя к F лемму Брамбла — Гильберта, будем иметь

$$| G([I - \Pi] v) | \leqslant C \| F \|_{-k-1,\,T_0} | v |_{k+1,\,T_0},$$

где

$$\| F \|_{-k-1,\,T} = \sup_{w \in \mathscr{H}_2^{(k+1)}(T_0)} \left\{ \frac{| G([I - \Pi] w) |}{\| w \|_{k+1,\,T}} \right\}.$$

Из двойственности пространств $\mathscr{H}_2^{(r)}(T_0)$ и $\mathscr{H}_2^{(-r)}(T_0)$ следует, что для любого $u \in \mathscr{H}_2^{(r)}(T_0)$

$$\| u \|_{r,\,T_0} = \sup_{G \in \mathscr{H}_2^{(-r)}(T_0)} \left\{ \frac{| G(u) |}{\| G \|} \right\},$$

так что, в частности,

$$\| (I - \Pi) v \|_{r,\,T} = \sup_G \left\{ \frac{| G([I - \Pi] v) |}{\| G \|} \right\}.$$

Объединяя теперь эти результаты, получим

$$\| (I - \Pi) v \|_{r,\,T_0} \leqslant C | v |_{k+1,\,T_0} \sup_G \left\{ \frac{\| F \|}{\| G \|} \right\} \leqslant$$

$$\leqslant C | v |_{k+1,\,T} \sup_G \left\{ \sup_{w \in \mathscr{H}_2^{k+1}(T_0)} \left\{ \frac{| G([I - \Pi] w) |}{\| w \|_{k+1,\,T_0}} \right\} \frac{1}{\| G \|} \right\},$$

но

$$|G([I-\Pi]w)| \leqslant \|G\| \|I-\Pi\| \|w\|_{k+1, T_0},$$

откуда сразу следует утверждение леммы.

Этот результат может быть объединен с условием регулярности для получения оценки ошибки интерполяции в области R.

Теорема 5.4. *Предположим, что условие полноты справедливо для некоторого $k > 0$ и условие регулярности справедливо для всех $r \leqslant k$. Тогда если $u \in \mathscr{H}_2^{(k+1)}(R)$, $\tilde{u}$ интерполирует u, то*

$$\|u-\tilde{u}\|_{r, R} \leqslant Ch^{k+1-r} |u|_{k+1, R},$$

где h ограничивает диаметры элементов, составляющих разбиение области R.

Доказательство. Если мы предположим, что область R разбита на элементы T_j $(j = 1, \ldots, S)$, то

$$\|u-\tilde{u}\|_{r, R}^2 = \sum_{j=1}^{S} \|u-\Pi_{T_j} u\|_{r, T_j}^2.$$

Рассматривая типичный элемент T, получим из леммы 5.5, что

$$\|u-\Pi u\|_{r, T} \leqslant C\|I-\Pi\| |u|_{k+1, T}.$$

Применяя затем условие 5.1 и результат упражнения 4, будем иметь

$$\|u-\Pi_T u\|_{r, T} \leqslant Ch^{k+1-r}\|I-\Pi\| |u|_{k+1, T}.$$

Ввиду равномерной ограниченности всех операторных норм $\|I-\Pi\|$ суммирование по всем элементам приводит к желаемому результату.

Этот результат может быть применен в том специальном случае, когда преобразование T_0 в T линейно и интерполяция конечноэлементными функциями точна для $u \in P_k$. Тогда, в частности,

$$\|u-\tilde{u}\|_{1, R} \leqslant Ch^k |u|_{k+1, R},$$

где $\tilde{u}$ — интерполирующая функция. Такой тип оценки требует для задач второго порядка определения порядка конечноэлементной аппроксимации с помощью почти оптимальных неравенств параграфа 5.2. Отметим, что если рассматривать область только из одного элемента, то получается оценка ошибки для обычной интерполяции Лагранжа; это обстоя-

тельство будет использовано ниже, в разд. (5.4(B)). Лагранжевы (или эрмитовы) элементы из разд. 4.1 таковы, что преобразование T_0 в T линейно, базисные функции $\varphi \in P_k$, интерполяция точна для $u \in P_k$ и $K_0 = K_{[T]} = P_k$. Поэтому если область R является многоугольником, граничные условия точно согласованы и интегралы вычислены аналитически, то для ошибки такой конечноэлементной аппроксимации задачи второго порядка будем иметь

$$\| u - U \|_{1, R} \leqslant Ch^k | u |_{k+1, R}. \tag{5.18}$$

Аналогичный результат справедлив для задач на прямоугольниках с прямоугольной сеткой: снова показатель степени точно интерполируемых полиномов определяет показатель степени у h в оценке (5.18).

Упражнение 16. Покажите, что исключение внутренних параметров из таких конечноэлементных аппроксимаций, как лагранжевы (или эрмитовы) кубические элементы, или исключение нормальных производных в серединах сторон 21-параметрической аппроксимации полиномами пятой степени со сшивкой в C^1 уменьшает показатель степени у h в (5.18) на единицу.

Упражнение 17. Покажите, что биквадратичная и бикубическая субпараметрические аппроксимации (разд. 4.3), использующие билинейное преобразование для перевода произвольного четырехугольника в единичный квадрат, могут интерполировать точно полиномы второй и третьей степени соответственно по x и y. Покажите, что если R является многоугольником, граничные условия точно согласованы и интегралы вычисляются аналитически, то оценка (5.18) справедлива при $k = 2$ и 3 соответственно.

Упражнение 18. Покажите, что оценка (5.18) справедлива для шестигранных субпараметрических элементов в трехмерном пространстве при условии, что область R допускает точное разбиение.

Криволинейные элементы

Сьярле и Равьяр (1972b) показали, что оценки вида (5.18) справедливы также для некоторых изопараметрических элементов с одной криволинейной стороной; в частности, они установили порядок сходимости для двух специальных случаев:

(1) *Квадратичные изопараметрические треугольные элементы* могут быть представлены в виде (упражнение 9)

$$t = t_3 + (t_1 - t_3)\,p + (t_2 - t_3)\,q + \left(t_4 - \frac{t_1 + t_2}{2}\right) 4pq \quad (t = x,\ y).$$

Так как интерполяция точна для полиномов второй степени по p и q, то ошибка оценивается как

$$\|u - U\|_{1,\,R} = O(h^2)$$

при условии, что

$$\left\{\left[x_4 - \frac{x_1 + x_2}{2}\right]^2 + \left[y_4 - \frac{y_1 + y_2}{2}\right]^2\right\}^{1/2} = O(h^2). \tag{5.19}$$

Это эквивалентно тому, что

$$\|P_4 - P_4^*\|_{\mathbb{R}^2} = O(h^2),$$

где P_4^* есть середина хорды P_1P_2, а норма — это евклидово расстояние в $\mathbb{R}^2$. Этому дополнительному условию всегда можно удовлетворить, если h достаточно мало по сравнению с радиусом кривизны границы. Предполагается также, что граница может быть точно представлена с помощью дуг, допускающих параметризацию вида

$$t = t_1 p\,(2p - 1) + t_2\,(1 - p)\,(1 - 2p) + t_4 4p\,(1 - p) \tag{5.20}$$
$$(t = x,\ y;\ p \in [0,\ 1]).$$

Аналогичная оценка справедлива для биквадратичных изопараметрических аппроксимаций, определенных на четырехугольниках с одной криволинейной стороной, если выполнены такие же ограничения геометрического характера.

(2) *Кубические изопараметрические элементы* могут быть представлены в виде (упражнение 10)

$$t = t_3 + (t_1 - t_3)\,p + (t_2 - t_3)\,q + 27pq\,(1 - p - q)\left(t_{10} - \frac{t_1 + t_2 + t_3}{3}\right) +$$
$$+ \frac{9}{2}\,pq\left[(3p - 1)\left(t_4 - \frac{2t_1 + t_2}{3}\right) + (3q - 1)\left(t_5 - \frac{2t_2 + t_1}{3}\right)\right]$$
$$(t = x,\ y).$$

Так как интерполяция точна для кубических полиномов по p и q, то ошибка оценивается как

$$\|u - U\|_{1,\,R} = O(h^3)$$

при условии, что

$$\|P_j - P_j^*\|_{\mathbb{R}^2} = O(h^2) \quad (j = 4,\ 5)$$

и

$$\|(P_4 - P_4^*) - (P_5 - P_5^*)\|_{\mathbb{R}^2} = O(h^3), \tag{5.21}$$

и точка P_{10} выбрана так, что

$$t_{10} = \frac{t_1 + t_2 + t_3}{3} + \frac{(t_4 - t_4^*) + (t_5 - t_5^*)}{4} \quad (t = x, y), \quad (5.22)$$

де $P_4^* = (x_4^*, y_4^*)$ и $P_5^*(x_5^*, y_5^*)$ — точки, делящие хорду P_1P_2 на три равные части и прилежащие к P_4 и P_5 соответственно. Дополнительное ограничение (5.21) в действительности выполняется при достаточно малом h. Как и в предыдущем случае, оценка порождаемой конечноэлементной аппроксимацией ошибки справедлива лишь тогда, когда в *граничном условии нет погрешности.* Подобная же оценка имеет место для эрмитовой изопараметрической аппроксимации (Сьярле и Равьяр, 1972b) при выполнении совокупности условий, аналогичных (5.21) и (5.22).

Эти строгие ограничения на кривизну изопараметрических элементов могут оказаться необходимыми на практике, как показывают некоторые численные эксперименты (Бонд, Свенелл, Геншелл и Уабартон, 1973), но комментируются они по-разному.

Во всех оценках сходимости для случая двух переменных предполагается, что наименьший угол θ треугольной сетки остается отделенным от нуля при стремлении h к нулю. В аналогичных результатах для случая трех переменных или четырехугольных сеток на плоскости предполагается, что отношение $\frac{h}{\rho}$ остается ограниченным при стремлении h к нулю. Для любого элемента величина ρ определяется как диаметр наибольшей сферы (в $\mathbb{R}^3$) или окружности (в $\mathbb{R}^2$), лежащей внутри этого элемента. Если же таких предположений нельзя сделать, то оценка для ошибки интерполяции примет вид

$$\| u - \tilde{u} \|_{r,R} = O\left(\frac{h^{k+1}}{\rho^r}\right) \quad (5.23)$$

в предположении, что величина $\|I - \Pi\|$ равномерно ограничена по всем элементам. Оценка такого вида вводится потому, что в условии 5.1 второе неравенство (5.7b) теперь правильнее выражать через ρ, чем через h (Сьярле и Равьяр, 1972a, Брамбл и Зламал, 1970). Для треугольных сеток

$$\rho \approx h \sin \theta,$$

и поэтому в скобках правой части (5.23) будет стоять $h^{k+1-r}/(\sin\theta)^r$. Если нельзя считать отношение $\frac{h}{\rho}$ ограничен-

ным при стремлении h к нулю, то может оказаться [1]), что величина $\|I - \Pi\|$ также не будет равномерно ограниченной, и поэтому, как показали Брамбл и Зламал, оценка для ошибки интерполяции может быть представлена в виде

$$\|u - \tilde{u}\|_{r,R} \leqslant C \frac{h^{k+1-r}}{(\sin\theta)^{n+r}} |u|_{k+1,R}$$

при некотором $n \geqslant 1$, что лучше оценки (5.18). Бабушка и Азиз (1976) отмечают, что лучше иметь дело с углами, стремящимися к 2π, чем с углами, стремящимися к нулю.

Оценки вида

$$\|u - \tilde{u}\|_{r,R} \leqslant C_{k,r}\, h^{k+1-r} \tag{5.24}$$

для ошибки интерполяции могут быть получены с помощью *теоремы Сарда о ядре* для методов, использующих как прямоугольники, так и треугольники. Для некоторых видов кусочных полиномиальных аппроксимаций значения постоянных $C_{k,r}$ из неравенства (5.24) могут быть получены численно. Результаты эти также подтверждают, что оценки вида (5.23) не являются наилучшими из возможных (см. Барнхилл, Грегори и Уайтман, 1972, и цитируемые ими работы).

Оценки ошибки интерполяции в терминах максимум-полунормы $|.|_{(k)\bar{R}}$ более правильные по сравнению с использующими соболевскую полунорму оценками вида (5.18), сначала получил Зламал, а затем Сьярле и Равьяр и Женишек (см. Сьярле и Равьяр, 1972а, и цитируемые ими работы).

Если решение не является достаточно гладким, величины $|u|_{k+1,R}$ могут не существовать и нельзя получить оценку вида (5.18), даже если $K_{[0]} \supset p_k$. В таких случаях необходимо использовать показатель

$$k^* = \max\{s : u \in \mathscr{H}_2^{(s)}(R)\};$$

так как $k^* \leqslant k$, то необходимо лишь переформулировать теорему 5.4, заменив k на k^* там, где это нужно; в остальном анализ проводится без изменений. Типичные задачи, для которых решение сингулярно или почти сингулярно, возникают в случае областей с вдавленными внутрь углами. Для решения таких задач были разработаны специальные приемы, которые кратко рассматриваются в разд. 7.4(F); другой подход приводится в гл. 8 книги Стренга и Фикса (1977) и в работе Уэйта (1976).

---

[1]) Одним из примеров такого рода является 21-параметрическая аппроксимация полиномами пятой степени со сшивкой элементов в C^1.

5.4. Ошибки возмущений

В оценках (5.16) и (5.17) содержатся два различных типа ошибок:

(1) Первый член

$$\inf_{V_h \in K_h} \|u - V_h\|$$

представляет собой *ошибку аппроксимации* и может быть оценен методами, изложенными в разд. 5.3.

(2) Все остальные члены возникают из-за приближенного характера уравнения Галеркина (5.12). В этом параграфе мы попытаемся оценить эти дополнительные члены для различных видов возмущений.

Конечноэлементное решение называется *оптимальным*, если для него порядок ошибок возмущений не превосходит порядка ошибки аппроксимации (Нитше, 1972).

В своем исследовании квадратурных ошибок Хеболд и Варга (1972) назвали квадратурную формулу *совместимой*, если в результате ее применения получается оптимальная аппроксимация.

(A) Численное интегрирование

Оценки ошибок интегрирования, осуществляемого численно с помощью стандартных квадратурных формул, были рассмотрены в упражнениях 6 и 7 разд. 5.1. В этом разделе мы рассмотрим этот вопрос более детально и получим оценки, справедливые при некоторых типах нелинейности в преобразовании, отображающем T_0 на T.

Квадратурная схема на стандартном элементе задается последовательностью точек $p_l \in T_0$ $(l = 1, \ldots, L)$ и последовательностью положительных весов b_l. Если возмущенная билинейная форма эллиптична в K_h, то условие $b_l > 0$ необходимо. Любой интеграл по стандартному элементу можно записать в виде

$$\iint_{T_0} u(\mathbf{p})\, d(\mathbf{p}) = \sum_{l=1}^{L} b_l u(\mathbf{p}_l) + E_0(u),$$

где, как и раньше, E_0 есть оператор квадратурной ошибки для стандартного элемента. Чтобы получить квадратурную формулу для интегрирования на произвольном элементе, возьмем точки $\mathbf{x}_l = \mathbf{F}(\mathbf{p}_l) \in T$ и веса $\beta_l = b_l J(\mathbf{p}_l)$, где J есть якобиан преобразования $\mathbf{F}$. Если обозначить через $E(u)$ квадра-

турную ошибку на элементе T, то

$$E(u) = \iint_T u(\mathbf{x})\, d\mathbf{x} - \sum_{l=1}^{L} \beta_l u(\mathbf{x}_l) =$$

$$= \iint_T u(\mathbf{p}) J(\mathbf{p})\, d\mathbf{p} - \sum_{l=1}^{L} b_l J(\mathbf{p}_l)\, u(\mathbf{p}_l) = E_0(uJ).$$

В этом разделе, следуя Сьярле и Равьяру (1972c), мы изучим ошибки возмущений для решения дифференциального уравнения

$$\frac{\partial}{\partial x}\left(d_1 \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(d_2 \frac{\partial u}{\partial y}\right) + f(x, y) = 0 \quad ((x, y) \in R) \quad (5.25)$$

с граничным условием

$$u(x, y) = 0.$$

Предположим, что все ошибки возмущений возникают при вычислении скалярных произведений с помощью численного интегрирования. Поэтому все базисные функции φ удовлетворяют граничному условию и согласованы — для этой задачи $\varphi \in \overset{\circ}{\mathscr{H}}{}_2^{(1)}(R)$. Предположим далее, что область R разбита на S элементов T_j $(j = 1, \ldots, S)$ и что для каждого элемента существует преобразование $\mathbf{F}_j$, отображающее на него стандартный элемент; якобиан преобразования $\mathbf{F}_j$ обозначим через J_j.

Поэтому нашей задаче будет соответствовать билинейная форма

$$a(u, v) = \sum_{j=1}^{S} \iint_{T_j} \left\{ d_1 \left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial v}{\partial x}\right) + d_2 \left(\frac{\partial u}{\partial y}\right)\left(\frac{\partial v}{\partial y}\right) \right\} dx\, dy,$$

и возмущенная форма будет иметь вид

$$a_h(u, v) = \sum_{j=1}^{S} \sum_{l=1}^{L} \beta_{lj} \left\{ d_1 \left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial v}{\partial x}\right) + d_2 \left(\frac{\partial u}{\partial y}\right)\left(\frac{\partial v}{\partial y}\right) \right\}_{\mathbf{x} = \mathbf{F}_j(\mathbf{p}_l)},$$

где

$$\beta_{lj} = b_l J_j(\mathbf{p}_l) \begin{cases} j = 1, \ldots, S \\ l = 1, \ldots, L. \end{cases}$$

Следовательно,

$$(a - a_h)(u, v) = \sum_{j=1}^{S} \left\{ E_0\left(d_1 \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} J_j\right) + E_0\left(d_2 \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} J_j\right) \right\}. \quad (5.26)$$

Таким образом, отдельные члены последней суммы имеют вид $E_0(zw)$, где $z(\mathbf{p})$ и $w(\mathbf{p})$ есть соответственно $d_1\left(\frac{\partial u}{\partial x}\right)$ и $\left(\frac{\partial v}{\partial x}\right)J_j$ (или $d_2\left(\frac{\partial u}{\partial y}\right)$ и $\left(\frac{\partial v}{\partial y}\right)J_j$); пусть d_1 и d_2 являются настолько гладкими, что, $z \in \mathscr{H}_2^{(k)}(T_0)$. Тогда для интегралов от произведений можно использовать оценку, полученную в упражнении 7, при условии, что $\left(\frac{\partial v}{\partial x}\right)J_j$ и $\left(\frac{\partial v}{\partial y}\right)J_j$ являются полиномами по $\mathbf{p}$.

Аналогичным образом квадратурная формула применяется и к правой части уравнения Галеркина

$$a(u, v) = (f, v),$$

где

$$(f, v) = \sum_{j=1}^{S} \iint_{T_j} f(\mathbf{x})\, v(\mathbf{x})\, d(\mathbf{x}).$$

Отсюда следует, что

$$(f, v)_h = \sum_{j=1}^{S} \sum_{l=1}^{L} \beta_{lj} \{f(\mathbf{x})\, v(\mathbf{x})\}_{\mathbf{x}=\mathbf{F}_j(\mathbf{p}_l)},$$

и поэтому

$$(f, v) - (f, v)_h = \sum_{j=1}^{S} E_0(fvJ_j). \tag{5.27}$$

Если снова $f \in \mathscr{H}_2^{(k)}(T_0)$ при некотором k, то для произведения такого вида можно применить оценки квадратурных ошибок при условии, что vJ_j есть полином по $\mathbf{p}$. Вместе с тем мы получили оценки возмущений (5.26) и (5.27), которые могут быть использованы в (5.16) и (5.17) при выяснении вопроса о сходимости аппроксимации.

Если каждый элемент получается из стандартного элемента линейным преобразованием, то все якобианы являются константами, и если $W_h(\mathbf{x})$ есть полином, то полиномами будут и $\frac{\partial W_h(\mathbf{p})}{\partial x}$, $\frac{\partial W_h(\mathbf{p})}{\partial y}$ и $W_h(\mathbf{p})$. Можно показать, что на каждом элементе функции $\partial W_h(\mathbf{p})/\partial x J(\mathbf{p})$, $\partial W_h(\mathbf{p})/\partial y J(\mathbf{p})$ и $W_h(\mathbf{p})J(\mathbf{p})$ будут полиномами, если $W_h(\mathbf{x})$ есть пробная конечноэлементная функция различных видов (Сьярле и Равьяр, 1972c).

Упражнение 19. Покажите, что для квадратичного изопараметрического элемента, приведенного в упражнении 9, функции $J(\partial p/\partial x)$, $J(\partial p/\partial y)$, $J(\partial q/\partial x)$ и $J(\partial q/\partial y)$ *линейны по* p и q. Покажите далее, что для любой кусочно-квадратичной пробной функции W_h на каждом треугольнике функции

$J(\mathbf{p})\partial W_h(\mathbf{p})/\partial x$ и $J(\mathbf{p})\partial W_h(\mathbf{p})/\partial y$ будут полиномами второй степени по $\mathbf{p}$; покажите также, что $J(\mathbf{p})W_h(\mathbf{p})$ есть полином четвертой степени.

Упражнение 20. Докажите, что если используются треугольные изопараметрические элементы степени k, то на каждом треугольнике функции J, $J(\partial w_h/\partial x)$ и $J(\partial W_h/\partial y)$ будут полиномами степени $2(k-1)$ для любой пробной функции W_h. Докажите, что в общем случае $JW_h \in P_{3k-2}$ на каждом треугольнике.

Сделав предположение о полиномиальности $J(\mathbf{p})\partial W_h/\partial x$, $J(\mathbf{p})\partial W_h/\partial y$ и $J(\mathbf{p})W_h$ на каждом элементе, можно получить оценки возмущений (5.26) и (5.27).

Теорема 5.5. *Предположим, что для любой пробной функции $W_h \in K_h$ на каждом элементе функции $J(\mathbf{p})\partial W_h/\partial x$ и $J(\mathbf{p})\partial W_h/\partial y$ есть полиномы степени не выше r_1, а $J(\mathbf{p})W_h$ — полином степени не выше r_0, и что условие регулярности выполняется для всех $s \leqslant \max\{r_1, r_0\}$. Тогда если квадратурная формула точна на стандартном треугольнике для полиномов степени не выше $r_1 + s$, то возмущение* (5.26) *билинейной формы ограничено как*

$$\frac{|(a-a_h)(V_h, W_h)|}{\|W_h\|_{1,R}} \leqslant Ch^{s+1}\|V\|_{s+2,R}, \tag{5.28}$$

где V_h, $W_h \in K_h$ есть любые пробные функции. Аналогично этому, если квадратура точна для всех полиномов степени не выше $r_0 + s - 1$, то возмущение (5.27) *правой части ограничено как*

$$\frac{|(f, W_h) - (f, W_h)_h|}{\|W_h\|_{1,R}} \leqslant Ch^{s+1}\|f\|_{s+1,R} \tag{5.29}$$

для любой $W_h \in K_h$ при условии, что $f \in \mathscr{H}_2^{(s+1)}(R)$.

Если к примеру, применяются *лагранжевы или эрмитовы элементы степени k*, преобразование T в T_0 линейно и на каждом элементе $J_0(\partial W_h/\partial x) \in P_{k-1}$, $J_0(\partial W_h/\partial y) \in P_{k-1}$ и $J_0 W_h \in P_k$. Поэтому $r_1 = k-1$, $r_0 = k$, и если используется квадратура, которая *точна для всех полиномов степени не выше* $2k-2$, то в (5.28) и в (5.29) $s = k-1$. Тогда из (5.17) следует, что

$$\|u - U_h\|_{1,R} \leqslant C\|u - \tilde{u}\|_{1,R} + \sup_{W_h \in K_h}\left\{\frac{|(a-a_h)(\tilde{u}, W_h)|}{\|W_h\|_{1,R}}\right\} + \sup_{W_h \in K_h}\left\{\frac{|(f, W_h) - (f, W_h)_h|}{\|W_h\|_{1,R}}\right\},$$

где $\tilde{u}$ интерполирует u. Как было показано в разд. 5.3 (следствие из теоремы 5.4),

$$\| u - \tilde{u} \|_{1,R} \leqslant Ch^k | u |_{k+1,R},$$

поэтому аппроксимация будет *оптимальной* и

$$\| u - U_h \|_{1,R} = O(h^k),$$

так как добавочные члены (5.28) и (5.29) есть также $O(h^k)$.

Эта теорема не только указывает степень точности квадратурной формулы, обеспечивающую оптимальность аппроксимации, но указывает также, что минимальная степень, обеспечивающая сходимость при стремлении h к нулю, есть $\min\{r_1, r_0 - 1\}$. В предыдущем примере она равнялась $k - 1$.

Доказательство теоремы 5.5. Ошибка в билинейной форме складывается из членов вида

$$E_0\left(d_1 \frac{\partial V_h}{\partial x} \frac{\partial W_h}{\partial x} J\right) = E_0(v, w),$$

где $v(\mathbf{p}) = d_1(\partial V_h/\partial x)$ и $w(\mathbf{p}) = J(\partial W_h/\partial x)$. Если $w \in P_{r_1}$ и квадратура точна для полиномов степени $r_1 + s$, то из леммы Брамбла — Гильберта следует, что (см. упражнение 7)

$$| E_0(vw) | \leqslant C \| w \|_{\mathscr{L}_2(T_0)} | v |_{s+1,T_0}.$$

Считая коэффициент d_1 достаточно гладким, получим с помощью (5.7b) и результата упражнения 5, что

$$| E_0(vw) | \leqslant Ch^{s+1} \left\| \frac{\partial W_h}{\partial x} \right\|_{\mathscr{L}_2(T)} \left\| \frac{\partial V_h}{\partial x} \right\|_{s+1,T} \leqslant Ch^{s+1} \| W_h \|_{1,T} \| V_h \|_{s+2,T}.$$

Просуммировав по всем элементам и поделив на $| W_h |_{1,R}$, получим (5.28).

Справедливость (5.29) установим, следуя Сьярле и Равьяру (1972c). Введем проекцию Π_0 на одномерное подпространство P_0 функций — констант на элементе T_0; так что для любой функции $u \in \mathscr{L}_2(T_0)$

$$\iint_{T_0} (u - \Pi_0 u)\, d\mathbf{p} = 0.$$

Типичный член в ошибке для правой части уравнений Галеркина может быть записан как

$$E_0(fW_h J) = E_0(fJ[I - \Pi_0] W_h) + E_0(fJ[\Pi_0] W_h).$$

Так как $J[I - \Pi_0] W_h \in P_{r_0}$, то из леммы Брамбла — Гильберта следует, что (см. упражнение 7)

$$| E_0(fJ[I - \Pi_0] W_h) | \leqslant C \| J[I - \Pi_0] W_h \|_{\mathscr{L}_2(T_0)} | f |_{s,T_0}.$$

Но Π_0 есть проекционный оператор, так что можно применить лемму 5.5, чтобы получить неравенство

$$\| [I - \Pi_0] W_h \|_{\mathscr{L}_2(T)} \leqslant C | W_h |_{1, T_0}$$

Поэтому из (5.7а) следует, что

$$| E_0(fJ [I - \Pi_0] W_h) | \leqslant C \{ \sup_{\mathbf{p} \in T} J(\mathbf{p}) \} | f |_{s, T} | W_h |_{1, T} \leqslant$$

$$\leqslant C h^{s+1} \| f \|_{s, T} \| W_h \|_{1, T}. \quad (5.30)$$

Аналогично этому, если $J(\mathbf{p}) \in P_r$ и $\Pi_0 W_h$ есть константа, то

$$| E_0(fJ [\Pi_0 W_h]) | \leqslant C \| J [\Pi_0 W_h] \|_{\mathscr{L}_2(T)} | f |_{s+1, T_0} \leqslant$$

$$\leqslant C \{ \sup_{\mathbf{p} \in (T_0)} J(\mathbf{p}) \} \| W_h \|_{\mathscr{L}_2(T_0)} | f |_{s+1, T} \leqslant$$

$$\leqslant C h^{s+1} \| f \|_{s+1, T} \| W_h \|_{\mathscr{L}_2(T)}. \quad (5.31)$$

Объединяя (5.30) с (5.31), суммируя по всем элементам и деля на $\| W_h \|_{1, R}$, получим желаемый результат.

Упражнение 21. Рассматривая ошибки вида

$$E_0 \left(\frac{\partial^2 V_h}{\partial x^2} \frac{\partial^2 W_h}{\partial x^2} J \right),$$

покажите, что если квадратура точна для полиномов степени $r_1 + s$ и произвольный элемент переводится в стандартный элемент *линейным* преобразованием, то возмущение $(a - a_h)$ для задач четвертого порядка допускает оценку

$$\frac{| (a - a_h)(V_h, W_h) |}{\| W_h \|_{2, R}} \leqslant C h^{s+1} \| V_h \|_{s+3, R}$$

при условии, что пробные функции согласованы и являются полиномами степени не выше чем $r_1 + 2$ на каждом элементе.

Результаты, аналогичные теореме 5.5, были получены Фиксом (1972) при изучении влияния квадратурных формул на лагранжеву и эрмитову конечноэлементные аппроксимации на многоугольной области. Квадратурные формулы исследовались также Хеболдом и Варгой (1972), но только для прямоугольных областей и в предположении, что билинейная форма a_h проинтегрирована точно.

Сьярле и Равьяр (1972с) применили результаты теоремы 5.5 к изопараметрическим аппроксимациям, определенным как на треугольниках, так и на четырехугольниках. Они показали также, как выбрать такую квадратурную формулу, чтобы билинейная форма a_h была эллиптична в K_h, и тем самым можно было обосновать применение оценки (5.17).

Однако эти результаты приводят к содержательным оценкам только тогда, когда используемые изопараметрические элементы имеют не более *одной криволинейной стороны.* Даже в этих случаях результаты определяются условиями, отмеченными в разд. 5.3. Большинство оценок квадратурных ошибок обобщается на случай $\mathbb{R}^m (m > 2)$ для дифференциального уравнения

$$\sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(d_{i,j}\frac{\partial u}{\partial x_j}\right) + f(\mathbf{x}) = 0,$$

где

$$\mathbf{x} = (x_1, \ldots, x_m).$$

Упражнение 22. Используя результаты упражнения 20, покажите, что изопараметрическая аппроксимация степени k оптимальна, если используется квадратурная формула, степень точности которой равна $4(k-1)$.

(B) Интерполируемые граничные условия

Предположим теперь, что аппроксимирующее подпространство K_h содержит функции, не обращающиеся в нуль на границе, но что интегралы вычисляются точно. Поэтому можно использовать оценку ошибки (5.15), в которой добавочным членом будет

$$\sup_{W_h \in K_h} \left\{ \frac{|(f, W_h) - a(u, W_h)|}{\|W_h\|} \right\}. \tag{5.32}$$

Рассмотрим сначала приближенное решение уравнения

$$a(u, v) = (f, v),$$

где

$$a(u, v) = \iint_R \left\{ \left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial v}{\partial x}\right) + \left(\frac{\partial u}{\partial y}\right)\left(\frac{\partial v}{\partial y}\right) \right\} dx\, dy,$$

что соответствует дифференциальному уравнению

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f(x, y) = 0 \quad ((x, y) \in R),$$

с граничным условием

$$u = 0 \quad ((x, y) \in \partial R).$$

Результаты этого раздела могут быть применены и в случае неоднородных граничных условий, если, как и в разд. 5.1, граничные значения допускают подходящее продолжение. Возможен и другой подход, использующий для оценки ошибок теорему 5.3.

Анализ методов, в которых граничные условия не удовлетворяются точно, почти всегда приводит к рассмотрению интегралов от граничных значений. Это относится также и к методам штрафов, описанным в разд. 5.4(D). Можно определить соболевское пространство $\mathscr{H}_2^{(k-1)}(\partial R)$ аналогично тому, как было определено пространство $\mathscr{H}_2^{(k-1)}(R)$ в разд. 5.1. Из такого определения следует, например, что

$$\left| \int_{\partial R} \left(\frac{\partial u}{\partial n} \right) W_h \, d\sigma \right| \leqslant \left\| \frac{\partial u}{\partial n} \right\|_{k-1,\,\partial R} \| W_h \|_{k-1,\,\partial R},$$

а в силу теоремы о следе

$$\left\| \frac{\partial u}{\partial n} \right\|_{k-1,\,\partial R} \leqslant C \| u \|_{k+1,\,R}.$$

Тогда из теоремы Грина следует, что

$$| (f, W_h) - a(u, W_h) | = \left| \int_{\partial R} \left(\frac{\partial u}{\partial n} \right) W_h \, d\sigma \right|,$$

и, объединив эти два выражения, мы получим оценку вида

$$| (f, W_h) - a(u, W_h) | \leqslant C \| u \|_{k+1,\,R} \| W_h \|_{1-k,\,\partial R}.$$

Чтобы быть применимым к задачам второго порядка, выражение (5.32) должно содержать $\| W_h \|_{1,\,R}$, но не должно содержать $\| W_h \|_{1-k,\,\partial R}$. По этой причине Скотт (1975) ввел оценки вида

$$\sup_{W_h \in K_h} \left\{ \frac{\| W_h \|_{1-k,\,\partial R}}{\| W_h \|_{1,\,R}} \right\} \leqslant C h^{k+1/2} \tag{5.33}$$

для пробных функций W_h, которые близки к нулю на границе ∂R. Бергер, Скотт и Стренг (1972) установили справедливость (5.33) для частного случая $k = 1$, но они не смогли указать наилучший способ выбора пробных функций и обобщить этот результат. Из определения двойственного пространства следует, что оценка (5.33) эквивалентна неравенству

$$\sup_{W_h \in K_h} \left\{ \sup_{g \in \mathscr{H}_2^{(k-1)}(\partial R)} \left\{ \frac{\left| \int_{\partial R} g W_h \, d\sigma \right|}{\| g \|_{k-1,\,\partial R} \| W_h \|_{1,\,R}} \right\} \right\} \leqslant C h^{k+1/2}$$

или тому, что при всех $g \in \mathscr{H}_2^{(k-1)}(\partial R)$ и $W_h \in K_h$

$$\left| \int_{\partial R} g W_h \, d\sigma \right| \leqslant C h^{k+1/2} \| g \|_{k-1,\,\partial R} \| W_h \|_{1,\,R}. \tag{5.34}$$

Лагранжевы элементы

Сейчас мы установим справедливость (5.34) для одного частного вида интерполируемых граничных условий, рассмотренного Скоттом; при этом мы будем в основном придерживаться его метода. Пусть область R разбита на треугольные элементы и стороны элементов прямолинейны внутри R, а примыкающие к границе элементы могут иметь одну криволинейную сторону. В этом разделе предполагается, что (криволинейная) граница для любого граничного элемента T_j может быть параметризована как

$$\partial R_j = \{(x_j(\theta),\ y_j(\theta)) : \theta \in [0,\ \Theta_j]\}.$$

С целью упрощения алгебраических выкладок предположим, что для типичного граничного элемента T координаты x и y выбраны так, что $\theta = x$ (рис. 25). Тогда для длины дуги $\sigma(x)$ будем иметь

$$\sigma(x) = \int_0^x \{1 + y'^2\}^{1/2}\, dx,$$

и

$$d\sigma = \frac{d\sigma(x)}{dx}\, dx.$$

Узлы интерполяции на искривленной стороне элемента определяются узлами $(k+1)$-точечной *квадратурной формулы Лобатто* на интервале $[0, 1]$, которые мы обозначим через $\xi_i^{[k]}$ $(i = 1, \ldots, k+1)$. Такая квадратурная формула рассматривается, например, в работе Дэвиса и Рабиновича (1967). Таким образом, для типичного элемента точками интерполяции будут $(\Theta\xi_i^{[k]},\ y(\Theta\xi_i^{[k]}))$ [1] $(i = 1, \ldots, k+1)$.

Таблица 3

Таблица узлов квадратурной формулы Лобатто на интервале [0,1]

$k=1$	$0, 1$
$k=2$	$0, \frac{1}{2}, 1$
$k=3$	$0, \frac{1}{2} \pm \frac{1}{2\sqrt{5}}, 1$
$k=4$	$0, \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{3}{7}}, \frac{1}{2}, 1$

[1]) Скотт показал, что это условие может быть несколько ослаблено без изменения окончательного результата.

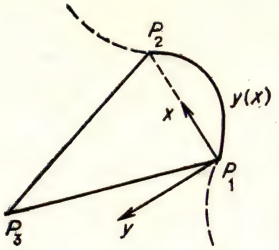


Рис. 25.

Интеграл от граничных значений из (5.34) можно переписать в виде

$$\int_{\partial R} gW_h\,d\sigma = \sum_j \int_{\partial R_j} gW_h\,d\sigma. \tag{5.35}$$

Типичное слагаемое этой суммы может быть представлено как

$$\int_0^\Theta z(x)\,w(x)\,dx, \tag{5.36}$$

где

$$z(x) = g(x,\ y(x))\,\frac{d\sigma}{dx}$$

и

$$w(x) = W_h(x,\ y(x)).$$

Теперь введем полином $\tilde{z}(x) \in P_{k-2}$, интерполирующий $z(x)$ на интервале $[0, \Theta]$. Из теоремы 5.4 следует (если заменить k на $k-2$ и $R = (0, \Theta) = I$), что

$$\|z - \tilde{z}\|_{r,\,I} \leqslant C\Theta^{k-r-1}\,|z|_{k-1,\,I} \quad (r \leqslant k-2), \tag{5.37}$$

если $z(x)$ и тем самым $g(x, y)$, и $\sigma(x)$ достаточно гладкие функции. Поэтому (5.36) можно представить как

$$\int_0^\Theta zw\,dx = \int_0^\Theta (z - \tilde{z})\,w\,dx + \int_0^\Theta \tilde{z}w\,dx. \tag{5.38}$$

Так как интервал $I = [0, \Theta]$ может быть переведен в стандартный интервал $[0, 1]$ линейным преобразованием, то из леммы Брамбла — Гильберта (см. упражнение 23) следует, что

$$\left|\int (z - \tilde{z})\,w\,dx\right| \leqslant C\Theta^{k-1}\,\|w\|_{\mathscr{L}_\infty(I)}\,|z|_{k-1,\,I}.$$

Но $W_h(x, y)$ обращается в нуль в $k+1$ граничных точках интерполяции, и поэтому $w(x)$ обращается в нуль в $k+1$ точ-

ках интервала I, так что по теореме Ролля

$$|w(x)| \leqslant C\Theta^{k+1} \left\{ \max_{[0,1]} \left| \frac{d^k w}{dx^k} \right| \right\}.$$

Поскольку Θ не может превосходить диаметра элемента T, из этого следует, что

$$|w(x)| \leqslant Ch^{k+1} \| W_h \|_{(k)\,\overline{T}}. \tag{5.39}$$

Можно показать (см. упражнение 24), что

$$h^k \| W_h \|_{(k)\,\overline{T}} \leqslant C \| W_h \|_{1,\,T}. \tag{5.40}$$

Объединяя два последних неравенства и интегрируя, будем иметь

$$\| w \|_{\mathscr{L}_2(I)} \leqslant Ch^{3/2} \| W_h \|_{1,\,T},$$

и поэтому

$$\left| \int (z - \tilde{z})\, w\, dx \right| \leqslant Ch^{k+1/2} \| g \|_{k-1,\,\partial R_T} \| W_h \|_{1,\,T}, \tag{5.41}$$

где ∂R_T обозначает криволинейную часть границы элемента T.

Рассмотрим теперь второе слагаемое в правой части (5.38). Так как функция $w(x)$ обращается в нуль в квадратурных точках Лобатто, интеграл, содержащий $w(x)$ как множитель в подынтегральном выражении, аппроксимируется нулем при использовании квадратуры Лобатто. Поэтому оценка квадратурной ошибки одновременно является и оценкой значения самого интеграла. Так как $(k+1)$-точечная квадратура Лобатто точна для полиномов степени не выше $2k-1$, из леммы Брамбла — Гильберта следует (см. упражнение 6), что

$$\left| \int_0^{\Theta} \tilde{z}(x)\, w(x)\, dx \right| \leqslant Ch^{2k} |\tilde{z}w|_{2k,\,I};$$

но $\tilde{z} \in P_{k-2}$ и $w(x) = W_h(x, y(x))$, где $W_h \in P_k$ на T, и поэтому

$$|\tilde{z}w|_{2k,\,I} \leqslant C \| \tilde{z} \|_{k-2,\,I} \| W_h \|_{k,\,I}.$$

Вместе с (5.37) это дает

$$\begin{aligned} \| \tilde{z} \|_{k-2,\,I} &\leqslant \| z \|_{k-2,\,I} + \| \tilde{z} - z \|_{k-2,\,I} \leqslant \\ &\leqslant C \| z \|_{k-1,\,I} \leqslant \\ &\leqslant C \| g \|_{k-1,\,\partial R_T}. \end{aligned}$$

Так как переход от T к стандартному элементу T_0 является линейным, то из условия регулярности будет следовать, что

$$\| W_h \|_{k,\,T} \leqslant Ch^{1-k} \| W_h \|_{1,T}.$$

Объединяя эти четыре последние неравенства, получим

$$\left| \int_0^{\Theta} \tilde{z} w \, dx \right| \leqslant Ch^{k+1} \| g \|_{k-1,\,\partial R_T} \| W_h \|_{1,\,T}. \tag{5.42}$$

Суммируя теперь неравенства вида (5.41) и (5.42) по всем прилегающим к границе элементам, получим для ошибки возмущения оценку

$$\left| \int_{\partial R} g W_h \, d\sigma \right| \leqslant Ch^{k+1/2} \| g \|_{k-1,\,\partial R} \| W_h \|_{1,\,R} \tag{5.43}$$

и, следовательно,

$$\left\{ \frac{| (f, W_h) - a(u, W_h) |}{\| W_h \|_{1,\,R}} \right\} \leqslant Ch^{k+1/2} \| u \|_{k+1,\,R}. \tag{5.44}$$

Теорема 5.4 показывает, что для интерполяционных полиномов Лагранжа степени k ошибка интерполяции оценивается как

$$\| u - \tilde{u} \|_{1,\,R} \leqslant Ch^k | u |_{k+1,\,R},$$

и поэтому оценка ошибки для аппроксимации Галеркина имеет вид

$$\| u - U_h \|_{1,\,R} \leqslant C \{ h^k | u |_{k+1,\,R} + h^{k+1/2} \| u \|_{k+1,\,R} \}. \tag{5.45}$$

Следовательно, конечноэлементная аппроксимация с такой интерполяцией граничных условий остается *оптимальной* до тех пор, пока ошибка возмущения будет более высокого порядка (по h), чем ошибка аппроксимации. Скотт (1975) и Чернука, Купер, Линдберг и Олсон (1972) предложили для треугольных элементов с криволинейными границами квадратурные формулы, сохраняющие порядок для кусочных квадратичных аппроксимаций. Такие аппроксимации изучались также Бергером (1973) с целью получения оценки ошибки в терминах нормы пространства $\mathscr{L}_2(R)$; он также проводил численную проверку порядков (1972). В противоположность интерполяции граничных данных по конечному числу значений можно строить аппроксимацию, точно воспроизводя их вдоль всей границы, если использовать смешанные функциональные интерполянты (Гордон и Уиксом, 1974). Некоторые сведения о смешанных функциях будут изложены в разд. 7.3.

Упражнение 23. Используя рассуждения, аналогичные приведенным в упражнении 6 (или в лемме 5.5), докажите, что из леммы Брамбла — Гильберта следует неравенство

$$\left|\int_0^\Theta (z-\tilde{z})\, w\, dx\right| \leqslant C\Theta^{k-1} \| w \|_{\mathscr{L}_2(I)} | z |_{k-1,\, I},$$

где $\tilde{z} \in P_{k-2}$ интерполирует z на интервале $I = (0, \Theta)$ и $w \in \mathscr{L}_2(I)$ (*Указание.* Для доказательства рассмотрите функционал $F \in \mathscr{L}\left(\mathscr{H}_2^{(k-1)}(I) \times \mathscr{L}_2(I);\ \mathbb{R}\right)$, где

$$F(z,\ w) = \int_0^\Theta (z-\tilde{z})\, w\, dx.)$$

Упражнение 24. Покажите, что для лагранжевых интерполяционных полиномов степени k на треугольниках с прямолинейными сторонами при $k = 2$ и $k = 3$ любая пробная функция W удовлетворяет неравенству

$$\| W \|_{(k)\,\overline{T}} \leqslant Ch^{-k} \| W \|_{1,\, T}.$$

Упражнение 25. Покажите, что оценка ошибки возмущения (5.44) справедлива также для частного вида эрмитовых кубических элементов, предложенных Скоттом (1975).

(C) Аппроксимация границы

По-видимому, первые оценки ошибок для методов конечных элементов на приближенно заданных областях были получены советскими математиками (см., например, работу Оганесяна, 1966). Они получили оценки для кусочно-линейных аппроксимаций на треугольных сетках и рассматривали только приближенное решение для задач второго порядка с граничным условием

$$\frac{\partial u}{\partial n} + \beta u = 0 \quad (\beta \geqslant 0),$$

заданным на криволинейной границе. Они показали, что

$$\| u - U_h \|_{1,\, R_h} \leqslant Ch \| u \|_{2,\, R_h},$$

но их доказательства слишком сложны и выходят за рамки этой книги. Для таких задач были получены оценки ошибки в терминах нормы пространства $\mathscr{L}_2(R)$ (Оганесян и Руховец, 1969). Несколько позже было показано (Стренг и Бергер, 1971, Томе, 1973), что если R_h есть многоугольник, вписанный в $R \subset \mathbb{R}^2$ ($\mathbb{R}^m$, $m \geqslant 2$ согласно Стренгу и Фиксу, 1977, пара-

граф 4.4), то для модельной задачи, определяемой уравнением

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f(x, y) = 0 \quad ((x, y) \in R) \tag{5.46}$$

и граничным условием $u = 0$ на ∂R, справедлива оценка

$$\| u - u_h \|_{1, R_h} = O(h^{3/2}),$$

где u_h есть решение возмущенной задачи, определяемой уравнением

$$\frac{\partial^2 u_h}{\partial x^2} + \frac{\partial^2 u_h}{\partial y^2} + f(x, y) = 0 \quad ((x, y) \in R_h) \tag{5.47}$$

и граничным условием $u_h = 0$ на ∂R_h.

Таким образом, если уравнение (5.46) решается приближенно путем разбиения многоугольной области R_h и последующего вычисления конечноэлементного решения уравнения (5.47), то, согласно результатам разд. 5.3,

$$\| u - U_h \|_{1, R_h} = O(h)$$

для кусочно-линейных аппроксимаций на треугольных разбиениях. Если аппроксимирующие функции содержат все полиномы степени 2 или выше, то, согласно тем же результатам,

$$\| u - U_h \|_{1, R_h} = O(h^{3/2}).$$

Этот порядок аппроксимации может оказаться существенно ниже того, который получился бы только на основании результатов разд. 5.3, и происходит это из-за плохой аппроксимации вблизи границы; иногда это понижение интерпретируют, как *эффект приграничного слоя*. Можно воспользоваться принципом максимума, чтобы показать, что при наличии негладкой границы возмущения будут меньше внутри R. Некоторые результаты такого рода могут быть распространены на тот случай, когда $R_h \not\subset R$. Свойства сходимости конечноэлементных аппроксимаций вовне области изучались Нитше и Шатцем (1974), а также Брамблом и Томе (1974).

Бергер, Скотт и Стренг (1972) показали, что если область R в общем случае аппроксимируется областью R_h, которая не обязательно является многоугольником, так, что максимальное расстояние между двумя границами ∂R и ∂R_h есть $O(h^{k+1})$, то соответствующий такому возмущению член в оценке (5.16) будет величиной порядка $O(h^{k+1/2})$. Этому условию может удовлетворять аппроксимация границы кусочными полиномами степени k, например, путем интерполяции; можно показать, что если для аппроксимации функций и интерполяции границы используются полиномы одинаковой степени, то

ошибка метода конечных элементов будет еще порядка $O(h^k)$ в терминах нормы пространства $\mathscr{H}_2^{(1)}(R_h)$. Аналогичный результат имеет место для изопараметрических аппроксимаций, так как Сьярле и Равьяр (1972c) показали, что заключения теоремы 5.5, будучи слегка измененными, остаются справедливыми и тогда, когда область задана приближенно. Похожий результат был получен Зламалом (1973, 1974). Задача Неймана рассматривалась несколькими авторами, например Стренгом и Фиксом (1973) и Бабушкой (1971).

(D) Методы штрафов

К этой категории относятся все те методы, в которых неоднородное граничное условие Дирихле рассматривается в форме интеграла от граничных значений, добавляемого к соответствующему функционалу, а не в форме наложения некоторого условия на аппроксимирующие функции. Такие методы могут основываться на методе наименьших квадратов или на методе Ритца, или же на сочетании их обоих. Наиболее часто употребляемый подход основывается на методе наименьших квадратов, для которого ошибки уже нельзя естественным образом получить в терминах соболевских норм, и приходится постоянно привлекать теорему о следе для оценки интегралов от граничных значений (см., например, гл. 6 в книге Варги, 1971).

Если мы предположим, что оценка ошибки интерполяции, приведенная в теореме 5.4, справедлива при некотором $k > 0$, т. е.

$$\| u - \tilde{u} \|_{r,\,R} \leqslant Ch^{k+1-r} \| u \|_{k+1,\,R},$$

то можно получить и непосредственную оценку ошибки, но не в терминах соболевских норм.

Теорема 5.6. *Если конечноэлементная аппроксимация удовлетворяет условию*

$$(AU_h - f,\ AV_h) = h^{-3} \langle U_h - g,\ V_h \rangle \quad (\textit{для всех } V_h \in K_h), \tag{5.48}$$

где

$$\langle \varphi,\ \psi \rangle = \int_{\partial R} \varphi \psi \, d\sigma,$$

и если теорема 5.4 *справедлива при некотором* $k > 0$, *то имеет место неравенство*

$$\| AU_h - Au \|_{\mathscr{L}_2(R)} + h^{3/2} \| U_h - u \|_{\mathscr{L}_2(\partial R)} \leqslant Ch^{k-1} \| u \|_{k+1,\,R}.$$

Доказательство. Так как

$$\| Au - A\tilde{u} \|_{\mathscr{L}_2(R)} \leqslant C \| u - \tilde{u} \|_{2,R}$$

для любой функции $u - \tilde{u} \in \mathscr{H}_2^{(2)}(R)$ и

$$\| u - \tilde{u} \|_{\mathscr{L}_2(\partial R)} \leqslant C \left\{ h^{-1/2} \| u - \tilde{u} \|_{\mathscr{L}_2(R)} + h^{1/2} \| u - \tilde{u} \|_{1,R} \right\}$$

(Агмон, 1965), то

$$\| Au - A\tilde{u} \|_{\mathscr{L}_2(R)} + h^{3/2} \| u - \tilde{u} \|_{\mathscr{L}_2(R)} \leqslant$$

$$\leqslant C \left\{ \| u - \tilde{u} \|_{2,R} + h^{-1} \| u - \tilde{u} \|_{1,R} + h^{-2} \| u - \tilde{u} \|_{\mathscr{L}_2(R)} \right\}. \quad (5.49)$$

Поэтому метод наименьших квадратов, заданный с помощью (5.48), является проекционным методом в том смысле, что

$$\| u - U_h \|_{[1]} = \inf_{\tilde{u} \in K_h} \| u - \tilde{u} \|_{[1]},$$

где норма определена как

$$\| \varphi \|_{[1]}^2 = \| A\varphi \|_{\mathscr{L}_2(R)}^2 + h^{-3} \| \varphi \|_{\mathscr{L}_2(\partial R)}^2.$$

Результат немедленно следует из неравенства (5.49) и теоремы 5.4 при $r = 0$, 1 и 2.

Можно доказать также, что

$$\| u - U_h \|_{\mathscr{L}_2(R)} \leqslant C h^{k+1} \| u \|_{k+1,R},$$

но доказательство выходит за рамки этой книги (Бейкер, 1973; или Брамбл и Шатц, 1970). Численный пример применения оценки такого частного вида будет приведен в разд. 7.4 (А); другие примеры можно найти у Сербина (1975).

Другие авторы предлагают отличные от изложенного проекционные методы для решения уравнения (5.46); они основывают свои методы на использовании таких норм, как

$$\| \varphi \|_{[2]}^2 = a(\varphi, \varphi) + h^{-1} \| \varphi \|_{\mathscr{L}_2(\partial R)}^2$$

(Брамбл, Дюпон и Томе, 1972) и

$$\| \varphi \|_{[3]}^2 = -a(\varphi, \varphi) - 2(A\varphi, \varphi) + h^2 \| A\varphi \|_{\mathscr{L}_2(R)}^2 +$$

$$+ \gamma \left\{ h^{-1} \| \varphi \|_{\mathscr{L}_2(\partial R)}^2 + h \left\| \frac{\partial \varphi}{\partial s} \right\|_{\mathscr{L}_2(\partial R)}^2 \right\} \quad (\gamma > 0)$$

(Брамбл и Нитше, 1973).

Такие методы допускают обобщение на задачи более высокого порядка и на задачи размерности большей двух. Пред-

лагаются также и методы, основанные на использовании стационарных точек функционалов, не являющихся положительно определенными (см., например, работу Томе, 1973). Методы штрафов изучались также Обэном (1972), с. 23.

(E) Несогласованные элементы

Определим билинейную форму

$$a_h(u, v) = \sum_{j=1}^{S} \iint_{T_j} \left\{ \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \right\} dx\, dy,$$

которая не совпадает с билинейной формой

$$a(u, v) = \iint_{R} \left\{ \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \right\} dx\, dy,$$

если функция v терпит разрыв вдоль границ между элементами; различие между двумя формами a и a_h является основным моментом при изучении несогласованных элементов (см. разд. 7.2); в дополнение к этому определим полунорму, соответствующую форме a_h, как

$$|u|_h = [a_h(u, u)]^{1/2},$$

а также норму

$$\|u\|_h = \left[|u|_h^2 + \|u\|_{\mathscr{L}_2(R)}^2 \right]^{1/2}.$$

В качестве примера результатов, которые могут быть получены для несогласованных элементов, рассмотрим пространство K_h кусочно-линейных элементов, которые согласованы в *серединах сторон* треугольной сетки (дальнейшие детали снова можно найти в параграфе 7.2). Для таких элементов ошибка при интерполяции линейных функций равна нулю, так что по аналогии с разд. 5.3

$$\inf_{V_h \in K_h} \|u - V_h\|_h \leqslant Ch\, |u|_{2,R}$$

для любой функции $u \in \mathscr{H}_2^{(2)}(R)$. Поэтому, если несогласованная аппроксимация оптимальна, добавочный член

$$\sup_{W_h \in K_h} \frac{|(f, W_h) - a_h(u, W_h)|}{\|W_h\|_h}$$

должен быть не больше, чем $O(h)$.

Применяя теорему Грина к каждому элементу, получим, что

$$(f, W_h) - a_h(u, W_h) = \sum_{j=1}^{S} \int_{\partial T_j} \frac{\partial u}{\partial n} W_h \, d\sigma =$$

$$= \sum_{l=1}^{Q} \int_{E_l} \left\{ \left(\frac{\partial u}{\partial n} W_h \right)^{[1]} + \left(\frac{\partial u}{\partial n} W_h \right)^{[2]} \right\} d\sigma,$$

где через E_l обозначена сторона между двумя соседними элементами сетки, а два слагаемых в последнем интеграле, отмеченные как [1] и [2] соответственно, есть предельные значения, соответствующие элементам по обе стороны линии разрыва E_l. Применяя лемму Брамбла — Гильберта к функционалам вида

$$F(u, W_h) = \int_{E_l} \left\{ \left(\frac{\partial u}{\partial n} W_h \right)^{[1]} + \left(\frac{\partial u}{\partial n} W_h \right)^{[2]} \right\} d\sigma \qquad (5.50)$$

(см. упражнение 27), можно показать (Крузей и Равьяр, 1973), что если

$$\int_{E_l} \{ W_h^{[1]} - W_h^{[2]} \} \, d\sigma = 0 \qquad (5.51)$$

для всех $W_h \in K_h$, то

$$|(f, W_h) - a_h(u, W_h)| \leqslant Ch |u|_{1,R} \|W_h\|_h. \qquad (5.52)$$

Упражнение 26. Покажите, что (5.51) выполняется для кусочно-линейных несогласованных элементов, описанных выше.

Отметим, что в разд. 7.2 будет показано, что (5.51) представляет собой кусочное тестирование несогласованных элементов для задач второго порядка. Сьярле (1973) получил аналогичные результаты для элементов, связанных с задачами об изгибе пластины. Сравнительный анализ несогласованных элементов для задач об изгибе пластины имеется у Ласко и Лесена (1975).

Упражнение 27. Предполагая выполненным условие (5.51), примените лемму Брамбла — Гильберта к функционалу (5.50) и докажите тем самым справедливость оценки (5.52).

5.5. Резюме

В этом разделе содержатся краткие выводы по некоторым из наиболее важных результатов разд. 5.2—5.4. [1])

[1]) Хотя этот раздел может читаться без знакомства с предыдущими параграфами данной главы, приведенные в нем результаты существенно используют обозначения и определения разд. 5.1.

Первый результат относится к интерполяционным свойствам базисных функций, используемых в конечноэлементной аппроксимации, и называется условием полноты. Если в общем случае полином степени k интерполируется точно и решение u задачи второго порядка, заданной на двумерной области R, является достаточно гладким, так что $u \in \mathscr{H}_2^{(k+1)}(R)$, то для аппроксимации Галеркина $U \in K_N$ справедлива оценка

$$\| u - U \|_{1,R} \leqslant C \| u - \tilde{u} \|_{1,R} \leqslant C h^k \| u \|_{k+1,R}, \tag{5.53}$$

где h есть диаметр наибольшего элемента и $\tilde{u} \in K_N$ интерполирует u. В этом контексте вид интерполирующей функции определяется типом используемой конечноэлементной аппроксимации; при этом предполагается, что все узловые параметры интерполяции определены точно.

Если решение u не является столь гладким, как этого требует оценка (5.53), то показатель степени у h в оценке ошибки уменьшается так, что при $u \in \mathscr{H}_2^{(k^*+1)}(R)$ $(k^* \leqslant k)$

$$\| u - U \|_{1,R} \leqslant C h^{k^*} \| u \|_{k^*+1,R}. \tag{5.54}$$

Оценка ошибки в такой модифицированной форме используется, например, тогда, когда какая-либо из младших производных имеет особенности внутри области R или на ее границе. Тот случай, когда $D^i u$ $(|i| \leqslant 1)$ имеет особенности на границе, подробно обсуждается в других работах (Стренг и Фикс, 1973, гл. 8, и Уэйт, 1976) и здесь рассматривается только вкратце в разд. 7.4(F).

Для аппроксимаций, построенных на элементах, у которых базисные функции не выражаются непосредственно через пространственные переменные x и y, а определяются вместо этого с помощью преобразования к стандартному элементу с новыми переменными p и q, показатель степени у h в (5.53) определяется немного иначе. Для таких аппроксимаций, в которые входят и весьма распространенные изопараметрические элементы, k есть степень полиномов по p и q, которые интерполируются точно.

При использовании криволинейных изопараметрических элементов важно помнить о тех строгих ограничениях, при которых справедливы оценки (5.53) и (5.54). Даже когда только одна сторона треугольного элемента заменяется отрезком квадратичной кривой, элемент будет близким к треугольному с прямолинейными сторонами только с точностью $O(h^2)$. Если криволинейная граница будет кубическим полиномом, то ограничения на элемент будут даже более строгими — детали изложены в разд. 5.3 на с. 132.

Если заданы граничные условия Дирихле, то необходимо добиться удовлетворения приближенным решением этих условий по всей длине границы, прежде чем применять оценки (5.53) или (5.54). Если же граничные условия интерполируются по конечному числу значений в граничных точках, оценка ошибки принимает вид

$$\| u - U \|_{1, R} \leqslant Ch^k \| u \|_{k+1, R} + \Delta, \tag{5.55}$$

где величина добавочного члена Δ определяется типом аппроксимации граничных данных. Оценкой (5.55) можно пользоваться также и тогда, когда область R аппроксимируется некоторым образом (например, многоугольником), когда применяется численное интегрирование или используются несогласованные элементы при построении аппроксимации. Все три приема можно рассматривать как отклонения от классического вариационного метода, и поэтому в анализе каждого из них есть много общего. Что же касается величины возмущения, то сходимость еще имеет место, если эта величина есть $O(h^s)$ $(s > 0)$. Измененная возмущением аппроксимация называется *оптимальной*, если порождаемая возмущением ошибка есть $O(h^k)$, т. е. имеет тот же порядок малости, что и ошибка интерполяции.

(A) Численное интегрирование

Можно показать, что для задач второго порядка, при решении которых используются элементы с прямолинейными сторонами, порождаемый численным интегрированием добавочный член есть

$$\Delta = O(h^s)$$

при условии, что квадратура точна для всех полиномов степени $s + k - 2$; это означает, что сходимость имеет место для квадратурной формулы порядка [1]) $k - 1$, а оптимальность — для квадратурной формулы порядка $2k - 2$. Если используются треугольные изопараметрические элементы с криволинейными сторонами, то в силу упомянутых выше ограничений для обеспечения оптимальности аппроксимации необходимы квадратурные формулы порядка $4(k - 1)$, тогда как для сходимости достаточно формул порядка $3(k - 1)$.

---

[1]) Квадратурная формула имеет порядок r, если она точна для всех полиномов степени не выше r.

(B) Интерполяция граничных условий

Если для каждого граничного элемента граничные условия согласованы только в конечном числе точек, то порядок соответствующего добавочного члена зависит от расположения этих точек интерполяции на границе. Аппроксимация будет оптимальной, если граничные данные согласованы в квадратурных точках Лобатто; это означает, что если граница элемента может быть задана параметрически как $(x(\theta), y(\theta))$, $(0 \leqslant \theta \leqslant \Theta)$, то точки границы, по значениям в которых интерполируются граничные данные, определяются квадратурными точками Лобатто на интервале $[0, \Theta]$. Другим способом задания квадратурных точек является использование длины дуги вдоль границы в качестве параметра.

(C) Аппроксимация криволинейных границ

Можно показать, что если для каждого граничного элемента граница аппроксимируется кривой, уравнение которой представляется полиномом степени k, то граница отстоит от своего приближения на $O(h^{k+1})$, и поэтому аппроксимация будет оптимальной. В противоположность этому, если границу аппроксимировать многоугольником, то соответствующий добавочный член будет величиной порядка $O(h^{3/2})$, и сохраняется только сходимость.

(D) Несогласованные элементы

Можно показать, что сходимость имеет место, если элементы выдерживают кусочное тестирование (см. гл. 7).

ГЛАВА 6

НЕСТАЦИОНАРНЫЕ ЗАДАЧИ

В гл. 3 мы построили семейство приближенных методов решения задач с граничными условиями; они сводятся к нахождению стационарной точки некоторого функционала, которая является также и точкой экстремума. В этой главе мы по возможности обобщим такие методы на задачи с начальными данными. Однако при рассмотрении вариационной формулировки эволюционных задач возникают дополнительные трудности. Например, в случае диссипативных систем после дополнения основной задачи сопряженной соответствующий им функционал $I(u, u^*)$ уже не будет обладать такими экстремальными свойствами. Даже в таких эволюционных задачах, для которых существует точная вариационная постановка, как, например, динамические системы Гамильтона, стационарная точка *не* является экстремальной.

Эти трудности привели различных авторов к предположению о том, что вариационные формулировки вряд ли окажутся полезными для решения нестационарных задач. Мы присоединяемся к этому мнению в особенности потому, что, как было показано в гл. 3, методом Галеркина можно пользоваться без какого-либо упоминания о вариационных принципах. Единственным оправданием изучения сопряженной задачи является желание рассмотреть диссипативные системы в рамках развитого здесь математического аппарата. Для подобных целей другие авторы предлагают так называемые ограниченные вариационные принципы или квазивариационные принципы; такие принципы не имеют большого внутреннего смысла, а просто служат математическим обоснованием для применения метода Галеркина к диссипативным системам. Все формулировки одинаково хороши в этом отношении и одинаково несовершенны в смысле строгости, когда дело касается задач с начальными данными.

6.1. Принцип Гамильтона

В разд. 2.5 уравнения движения непрерывно распределенной динамической системы были получены как необходимые условия стационарности функционала. Покажем теперь, что если используется этот подход, то приближенное решение мо-

жет быть получено, как и в гл. 3, путем определения стационарного значения относительно аппроксимирующего подпространства функций. Поскольку уравнения движения таких динамических систем будут гиперболического или параболического типа, соответствующий функционал не будет положительно определенным. Следовательно, даже для консервативных систем стационарное значение *не* является экстремумом и невозможно получить наилучшую аппроксимацию. Примером такого функционала, соответствующего уравнению движения

$$\frac{\partial^2 u}{\partial t^2} c^2 - \frac{\partial^2 u}{\partial x^2} = 0 \tag{6.1}$$

для колеблющейся струны, является выражение

$$I(v) = \frac{1}{2} \rho \int_{t_0}^{t_1} \int_{0}^{l} \left[\left(\frac{\partial v}{\partial t} \right)^2 - c^2 \left(\frac{\partial v}{\partial x} \right)^2 \right] dx\, dt. \tag{6.2}$$

Граничные условия

Может показаться, что приближенное решение вида

$$U(x, t) = \sum_{i=1}^{N} \alpha_i \varphi_i (x, t)$$

для уравнения (6.1) получается путем непосредственного применения изложенного в разд. 3.1 метода Ритца к функционалу $I(v)$, заданному формулой (6.2); к сожалению, это приводит к несовместности в граничных условиях.

Задача, определяемая линейным *гиперболическим* дифференциальным уравнением вида

$$\frac{\partial^2 u(\mathbf{x}, t)}{\partial t^2} + Au(\mathbf{x}, t) = f(\mathbf{x}, t) \quad (t_0 < t \leqslant t_1) \tag{6.3}$$

при $\mathbf{x} \in R \subset \mathbb{R}^m$ с границей ∂R, где A есть линейный эллиптический дифференциальный оператор, аналогичный введенному в гл. 3, корректно поставлена в области $R \times [t_0, t_1]$, если заданы граничные и начальные условия

$$u(\mathbf{x}, t) = 0 \quad ((\mathbf{x}, t) \in \partial R \times (t_0, t_1)), \tag{6.3a}$$

$$u(\mathbf{x}, t_0) = u_0(\mathbf{x}) \quad (\mathbf{x} \in R) \tag{6.3b}$$

и

$$\frac{\partial u(\mathbf{x}, t_0)}{\partial t} = v_0(\mathbf{x}) \quad (\mathbf{x} \in R). \tag{6.3c}$$

Однако задача не будет корректной, если (6.3c) заменить условием при $t = t_1$, таким, например, как

$$u(\mathbf{x}, t_1) = u_1(\mathbf{x}) \quad (\mathbf{x} \in R). \tag{6.3d}$$

Задача, определяемая уравнением (6.3) и условиями (6.3a), (6.3b) и (6.3d), возникает при определении стационарного значения такого функционала

$$I(v;\ t_0,\ t_1) = \int_{t_0}^{t_1} \{(v_t,\ v_t) - a(v,\ v) + 2(f,\ v)\}\, dt, \qquad (6.4)$$

где скалярное произведение обозначает то же самое, что и в главе 3.

Таким образом, перед вычислением решения Ритца для задачи, определяемой уравнением (6.3) и условиями (6.3a), (6.3b) и (6.3c), необходимо несколько видоизменить метод так, чтобы избежать введения граничного условия (6.3d). Мы не встретимся с этой трудностью, если воспользуемся полудискретным методом Канторовича, чтобы получить приближенное решение вида

$$U(\mathbf{x},\ t) = \sum_{i=1}^{N} \alpha_i(t)\, \varphi_i(\mathbf{x}),$$

поскольку при этом уравнение (6.3) будет заменено системой обыкновенных дифференциальных уравнений второго порядка относительно функций $\alpha_i(t)$ $(i = 1, \ldots, N)$ (см. разд. 3.3). Условия (6.3b) и (6.3c) заменятся эквивалентными условиями

$$\alpha_i(t_0) = c_i \quad (i = 1, \ldots, N) \qquad (6.5a)$$

и

$$\frac{d\alpha_i(t_0)}{dt} = d_i \quad (i = 1, \ldots, N). \qquad (6.5b)$$

Мы отложим обсуждение полудискретного метода до последнего параграфа и вернемся вкратце к задаче о вычислении решения Ритца вида

$$U(\mathbf{x},\ t) = \sum_{i=1}^{N} \alpha_i \varphi_i(\mathbf{x},\ t). \qquad (6.6)$$

Если функция $u(\mathbf{x}, t)$ является решением уравнения (6.3) и удовлетворяет заданным условиям, то при любых T_0 и T_1, таких, что $t_0 \leqslant T_0 < T_1 \leqslant t_1$, функция $u(\mathbf{x}, t)$ доставляет стационарное значение интегралу $I(v\colon T_0, T_1)$, где все допустимые функции удовлетворяют условиям

$$v(\mathbf{x},\ T_0) = u(\mathbf{x},\ T_0) \quad (\mathbf{x} \in R)$$

и

$$v(\mathbf{x},\ T_1) = u(\mathbf{x},\ T_1) \quad (\mathbf{x} \in R),$$

и граничному условию

$$v(\mathbf{x},\ t) = 0 \quad ((\mathbf{x},\ t) \in \partial R \times (T_0,\ T_1)).$$

В частности, если взять разбиение интервала $[t_0, t_1]$ точками

$$t_0 = \tau_0 < \tau_1 < \ldots < \tau_K = t_1,$$

то решение уравнения (6.3) доставит стационарное значение каждому из функционалов

$$I_n(v) = \int_{\tau_{n-1}}^{\tau_{n+1}} \{(v_t, v_t) - a(v, v) + 2(f, v)\}\, dt \quad (n = 1, \ldots, K). \tag{6.7}$$

Отметим, что области задания функционалов по времени перекрываются. Используя функционалы (6.7) вместо функционала (6.4), можно построить метод решения в виде последовательных шагов, сводящихся к методу Ритца. А именно, считая решение известным на момент $t = \tau_{n-1}$, попытаемся найти численно стационарную точку функционала I_n и тем самым решение на момент $t = \tau_{n+1}$, чтобы затем использовать его как известное условие. Чтобы осуществить это практически, представим задачу, так, как если бы граничное условие (6.3а) было задано вместе со значениями решения $u(\mathbf{x}, t)$ на моменты $t = \tau_{n-1}, \tau_{n+1}$, и затем *решим* ее относительно неизвестного решения на момент $t = \tau_{n+1}$. Такой подход возможен при условии, что у нас имеется дополнительная информация о решении на промежуточном шаге $t = \tau_n$. Следовательно, для вычисления решения на момент $t = \tau_{n+1}$ мы должны знать его на двух предыдущих шагах, т. е. при $t = \tau_n$ и $t = \tau_{n-1}$.

В качестве примера мы опишем один из способов вычисления конечноэлементного решения уравнения (6.1). Возьмем в области $0 \leqslant x \leqslant l$, $\tau_{n-1} \leqslant t \leqslant \tau_{n+1}$ разбиение вида

$$0 = x_0 < x_1 < \ldots < x_{L+1} = l,$$

где $x_{i+1} - x_i = \Delta x$ $(i = 0, 1, 2, \ldots, L)$, и будем также считать, что

$$\tau_{n+1} - \tau_n = \tau_n - \tau_{n-1} = \Delta t.$$

Теперь представим себе, что мы решаем краевую задачу, определяемую уравнением

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad (0 < x < l;\ \tau_{n-1} < t < \tau_{n+1}), \tag{6.8}$$

и условиями

$$u(x, \tau_{n+1}) = u_{n+1}(v) \quad (0 \leqslant x \leqslant l), \tag{6.9a}$$

$$u(x, \tau_{n-1}) = u_{n-1}(x) \quad (0 \leqslant x \leqslant l) \tag{6.9b}$$

и

$$u(0, t) = u(l, t) = 0 \quad (\tau_{n-1} < t < \tau_{n+1}). \tag{6.9c}$$

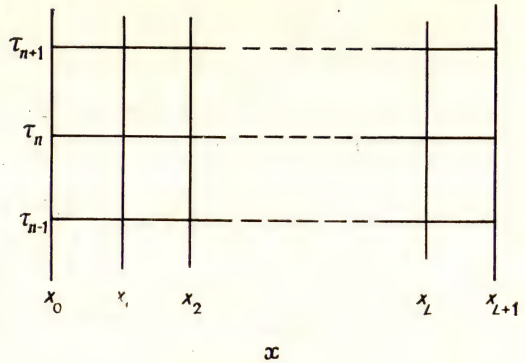


Рис. 26.

Чтобы получить приближенное решение этой задачи, определим в области $(0 \leqslant x \leqslant l;\ \tau_{n-1} \leqslant t \leqslant \tau_{n+1})$ билинейные базисные функции $\varphi_{ij}(x, t)$ $(i = 1, \ldots, L;\ j = n-1,\ n,\ n+1)$, соответствующие точкам $P_i^j = (x_i, \tau_j)$ (ср. с разд. 3.1), изображенным на рис. 26. Тогда приближенное решение запишется в виде

$$U(x, t) = \sum_{j=n-1}^{n+1} \sum_{i=1}^{L} \varphi_{ij}(x, t)\, U_i^j, \tag{6.10}$$

где U_i^j есть значение этого решения в точке P_i^j. При решении краевой задачи (6.8)—(6.9c) величины U_i^{n+1} и U_i^{n-1} $(i = 1, \ldots, L)$ определяются из (6.9a) и (6.9b) соответственно, а U_i^n определяются из системы

$$\frac{\partial}{\partial U_i^n} I_n \left(\sum_{j=n-1}^{n+1} \sum_{k=1}^{L} \varphi_{kj} U_k^j \right) = 0 \quad (i = 1, \ldots, L). \tag{6.11}$$

Мы же пытаемся решить не краевую задачу, а задачу с начальными значениями. Следовательно, если величины U_i^{n-1} и U_i^n $(i = 1, \ldots, L)$ известны, система (6.11) должна быть использована для определения U_i^{n+1}.

Упражнение 1. Покажите, что пошаговый метод решения уравнения

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0 \quad (0 < x \leqslant l;\ t_0 < t \leqslant t_1)$$

с использованием билинейных базисных функций, который был описан выше, приводит к системе разностных уравнений

$$\{\delta_t^2 I_x - r^2 \delta_x^2 I_t\} U_i^n = 0, \tag{6.12}$$

где δ_t^2 и δ_x^2 есть операторы взятия центральных разностей второго порядка, I_x и I_t — операторы интегрирования по Симпсону, введенные в разд. 3.1, а $r = \Delta t/\Delta x$, и область разбита так, что

$$\tau_{n+1} - \tau_n = \Delta t \quad (n = 0, \ldots, K) \text{ и } x_{i+1} - x_i = \Delta x \, (i = 0, \ldots, L).$$

Начальные условия

В приведенном выше примере и вообще в пошаговых методах, получающихся аналогичным образом для задачи (6.3)—(6.3c), необходимо как-то получить приближенное решение при $t = \tau_0$ и $t = \tau_1$, чтобы затем уже можно было использовать функционал $I_1(v)$ для нахождения решения при $t = \tau_2$. Если между начальными и граничными условиями нет разрывов, то приближенное решение можно определить в виде

$$U(\mathbf{x}, t) = u_0(\mathbf{x}) + \sum_{i=1}^{N} \alpha_i \varphi_i(\mathbf{x}, t). \tag{6.13}$$

Использование этого представления эквивалентно переходу к такой задаче (ср. с разд. 3.2), для которой начальное условие (6.3b) заменено условием

$$U(\mathbf{x}, t_0) = 0 \quad (\mathbf{x} \in R).$$

При наличии разрывов, т. е. когда

$$u_0(\mathbf{x}) \neq 0 \quad (\mathbf{x} \in \partial R),$$

нельзя использовать приближение вида (6.13) и необходимо аппроксимировать $u_0(\mathbf{x})$ функцией вида

$$U(\mathbf{x}, t_0) = \sum_{i=1}^{N} \alpha_i \varphi_i(\mathbf{x}, t_0)$$

так, что либо $U(\mathbf{x}, t_0)$ интерполирует $u_0(\mathbf{x})$, либо ошибка $u_0(\mathbf{x}) - U(\mathbf{x}, t_0)$ минимизируется в некоторой норме. Точно так же необходимо аппроксимировать и второе начальное условие, так что

$$\frac{\partial U(\mathbf{x}, t_0)}{\partial t} = \sum_{i=1}^{N} \alpha_i \frac{\partial \varphi_i(\mathbf{x}, t_0)}{\partial t}$$

является хорошим приближением для $v_0(\mathbf{x})$.

Поэтому в рассмотренном выше примере можно ввести начальные условия

$$U_i^0 = u_0(x_i) \quad (i = 1 \ldots, L)$$

и

$$\frac{U_i^1 - U_i^0}{\Delta t} = v_0(x_i) \quad (i = 1, \ldots, L).$$

Такое использование принципа Гамильтона позволяет построить вполне работоспособный метод последовательных шагов для решения консервативных систем, хотя метаматическая формулировка такого метода в ряде случаев не кажется слишком убедительной. Аналогичным образом пошаговый метод для приближенного решения гиперболических уравнений описывается у Нобла (1973).

6.2. Диссипативные системы

Как уже упоминалось в начале этой главы, различные авторы пытались получить вариационную формулировку для диссипативных задач. Некоторые из них пытались получить такой же общий вариационный принцип, который был бы справедлив для большого класса таких задач (Финлейсон и Скривен, 1967, и цитируемые ими работы), как принцип Гамильтона — для консервативных систем. Основные недостатки подобных формулировок таковы:

(1) Вариационные принципы проще всего получаются из основных уравнений, так что вариационная формулировка требует дополнительных усилий, но не дает дополнительной информации.

(2) Участвующий в формулировке вариационного принципа функционал не имеет физического смысла и его стационарные значения никогда не являются настоящими экстремумами, которые могли бы быть использованы для получения оценок ошибки.

(3) Как уже отмечалось в предыдущем разделе, имеется заметная несогласованность между вариационными задачами и задачами с начальными данными, которая постоянно игнорируется в большинстве так называемых вариационных формулировок эволюционных задач.

Мы подчеркнули достоинства и недостатки вариационных формулировок именно в этом месте книги ввиду большого разнообразия в литературе таких формулировок для *диссипативных* или *необратимых* задач и подчас противоречивой их трактовки. Мы не будем рассматривать конечноэлементное решение эволюционных диссипативных систем, получающееся исходя из сопряженной формы задачи, а вместо этого приведем два упражнения, которые могут быть выполнены интересующимся читателем.

Упражнение 2. Покажите, что при подходящем выборе аппроксимирующих функций функционал, соответствующий задаче, определяемой простейшим уравнением диффузии

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad (0 < x < l; \quad t_0 < t \leqslant t_1),$$

граничными условиями

$$u(0, t) = u(l, t)(x) = 0 \quad (t_0 < t \leqslant t_1)$$

и начальным условием

$$u(x, t_0) = u_0(x) \quad (0 \leqslant x \leqslant l),$$

имеет вид

$$I = \int_{t_0}^{t_1} \int_0^l \left\{ \frac{\partial v}{\partial t} v^* + \frac{\partial v}{\partial x} \frac{\partial v^*}{\partial x} \right\} dx\, dt.$$

Покажите далее, что уравнение

$$-\frac{\partial u^*}{\partial t} = \frac{\partial^2 u^*}{\partial x^2}$$

является необходимым условием для стационарности точки, определяемой из системы $\delta_v I(v, v^*) = 0$, если либо (I) функция u считается известной при $t = t_0$ и $t = t_1$, либо (II) u считается известной при $t = t_0$ и $u^* = 0$ при $t = t_1$.

Замечание. Отсюда видно, что при вычислении стационарной точки функционала I_n нужно предположить, что функция $u(x, \tau_{n-1})$ может принимать любые значения, тогда как $u^*(x, \tau_n) = 0$. Поэтому функция $U^*(x, t)$ — это не приближенное решение отдельной сопряженной системы, а скорее последовательность приближений для ряда отличных друг от друга сопряженных систем, соответствующих каждому из интервалов $[\tau_{n-1}, \tau_n]$, причем для каждой такой системы должно выполняться условие

$$U^*(x, \tau_n) = 0 \quad (n = 1, 2, \ldots, K)$$

при $0 \leqslant x \leqslant l$.

Упражнение 3. Покажите, что решая методом последовательных шагов уравнение

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \,(0 < x < l;\ t_0 < t \leqslant t_1),$$

как это было описано выше, и используя при этом билинейные базисные функции, мы придем к системе разностных уравнений

$$\Delta_t I_x U_i^n - \frac{1}{3} r \delta_x^2 \{U_i^{n+1} + 2U_i^n\} = 0, \tag{6.14}$$

где Δ_t есть оператор взятия разности вперед по t; I_x, δ_x^2 — определенные ранее разностные операторы по x, а $r = \Delta t/(\Delta x)^2$. Область разбита так, что $\tau_{n+1} - \tau_n = \Delta t$ $(n = 0, 1, \ldots, K)$ и $x_{i+1} - x_i = \Delta x$ $(i = 0, 1, \ldots, N)$. Здесь читатель, знакомый

с конечноразностными методами, может сравнить полученное выше разностное уравнение со стандартной конечноразностной аппроксимацией простейшего уравнения диффузии, такой, например, как схема Кранка — Николсона. Аналогичное сравнение можно провести между уравнением (6.12) и стандартными конечноразностными аппроксимациями волнового уравнения.

Другие «вариационные» приемы численного решения простейшего уравнения диффузии можно найти у Нобла (1973), а также в работе Чекки и Челла (1973).

6.3. Полудискретный метод Галеркина

Полудискретные методы, кратко упомянутые в разд. 6.1, позволяют обойтись без вариационной постановки эволюционных задач, включающей все независимые переменные. Они составляют основу наиболее употребительных методов решения таких задач. В гл. 3 различные модификации полудискретного метода применялись к эллиптическим задачам, но там такой подход оказывается несущественным. Первым шагом является переход к дифференциальному уравнению в слабой форме — как это было при рассмотрении методов Галеркина для эллиптических задач. Как и в гл. 3, здесь не делается попыток доказать эквивалентность классического и галеркинского решений дифференциального уравнения: будем считать, что решение единственно и совпадает с ними обоими.

В качестве примера рассмотрим дифференциальное уравнение

$$\frac{\partial u(\mathbf{x}, t)}{\partial t} + Au(\mathbf{x}, t) = f(\mathbf{x}, t) \quad (\mathbf{x}, t) \in R \times (t_0, t_1]), \tag{6.15}$$

где A есть дифференциальный оператор второго порядка, который в двумерном случае имеет вид

$$A = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}.$$

Уравнение (6.15) дополняется начальным условием

$$u(\mathbf{x}, t_0) = u_0(\mathbf{x}) \quad (\mathbf{x} \in R) \tag{6.15a}$$

и граничным условием

$$u(\mathbf{x}, t) = 0 \quad ((\mathbf{x}, t) \in \partial R \times (t_0, t_1]). \tag{6.15b}$$

Соответствующая слабая форма задачи в обозначениях параграфа 3.4 такова, что при $t \in (t_0, t_1]$

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u, v) = (f, v) \quad \text{(для всех } v(\mathbf{x}) \in \mathcal{H}) \tag{6.16}$$

совместно с начальным условием

$$(u, v)_{t=t_0} = (u_0, v) \quad (\text{для всех } v(\mathbf{x}) \in \mathscr{H}). \tag{6.16a}$$

Ясно, что для этой модельной задачи (если вспомнить обозначения гл. 5) $\mathscr{H} = \overset{\circ}{\mathscr{H}}{}_2^{(1)}(R)$ и решение $u \in \mathscr{H} \times C^1[t_0, t_1]$. Если заданы граничные условия более общего вида, то может оказаться необходимым модифицировать слабую форму задачи добавлением интегралов от граничных значений. Такая модификация функционалов рассматривалась в гл. 2, 3 и 5.

Полудискретная аппроксимация U теперь определяется из уравнения в слабой форме: это означает, что при $t \in (t_0, t_1]$ выполняется уравнение

$$\left(\frac{\partial U}{\partial t}, V\right) + a(U, V) = (f, V) \quad (\text{для всех } V(\mathbf{x}) \in K_N) \tag{6.17}$$

с начальным условием

$$(U, V)_{t=t_0} = (u_0, V) \quad (\text{для всех } V(\mathbf{x}) \in K_N). \tag{6.17a}$$

Для модельной задачи $K_N \subset \overset{\circ}{\mathscr{H}}{}_2^{(1)}(R)$. Если функции φ_i $(i = 1, \ldots, N)$ образуют базис подпространства K_N, то можно дать эквивалентную формулировку: полудискретная аппроксимация при $t \in (t_0, t_1]$ удовлетворяет уравнению

$$\left(\frac{\partial U}{\partial t}, \varphi_i\right) + a(U, \varphi_i) = (f, \varphi_i) \quad (i = 1, \ldots, N) \tag{6.18}$$

и условиям

$$(U, \varphi_i)_{t=t_0} = (u_0, \varphi_i) \quad (i = 1, \ldots, N). \tag{6.18a}$$

Для этой модельной задачи снова следует, что V и U должны обладать одинаковыми свойствами (по $\mathbf{x}$), так что $U \in K_N \times C^1[t_0, t_1]$, а аппроксимация Галеркина имеет вид

$$U(\mathbf{x}, t) = \sum_{i=1}^{N} \alpha_i(t)\, \varphi_i(\mathbf{x}). \tag{6.19}$$

Если заданы неоднородные граничные условия Дирихле, то аппроксимацию Галеркина можно представить в виде

$$U(\mathbf{x}, t) = W(\mathbf{x}, t) + \sum_{i=1}^{N} \alpha_i(t)\, \varphi_i(\mathbf{x}),$$

где $\varphi_i \in K_N$, а $W(\mathbf{x}, t)$ удовлетворяет граничным условиям. Как и для эллиптических задач, функция V должна только принадлежать энергетическому пространству, а на аппроксимацию U такого требования не накладывается.

Аппроксимация Галеркина определяется из системы обыкновенных дифференциальных уравнений относительно функ-

ций $\alpha_i(t)$ $(i = 1, \ldots, N)$. Из (6.18) следует, что для модельной задачи эти уравнения могут быть записаны как

$$\sum_{j=1}^{N} \left\{ \frac{d\alpha_j}{dt} (\varphi_j, \varphi_i) + \alpha_j a (\varphi_j, \varphi_i) \right\} = (f, \varphi_i) \; (i = 1, \ldots, N), \quad (6.20)$$

а начальные условия (6.18а) примут вид

$$\alpha_j (t_0) = c_j \quad (j = 1, \ldots, N), \quad (6.20a)$$

где

$$\sum_{j=1}^{N} c_j (\varphi_j, \varphi_i) = (u_0, \varphi_i) \quad (i = 1, \ldots, N). \quad (6.20b)$$

Определяемые из (6.20b) коэффициенты c_j $(j = 1, \ldots, N)$ удовлетворяют условию

$$\left\| u_0 (\mathbf{x}) - \sum_{j=1}^{N} c_j \varphi_j (\mathbf{x}) \right\|^2_{\mathscr{L}_2(R)} = \min;$$

для определенных задач это обстоятельство может быть использовано с целью замены (6.20b) другой аппроксимацией исходных данных или для изменения формы аппроксимации таким образом, чтобы точно удовлетворить начальному условию (разд. 6.1).

В качестве иллюстрации этого метода рассмотрим простейшее уравнение диффузии

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad (0 < x < l; t > 0),$$

решение которого удовлетворяет условиям

$$u(0, t) = u(l, t) = 0 \quad (t \geqslant 0)$$

и

$$u(x, 0) = u_0(x) \quad (0 < x < l).$$

Приближенное решение этой задачи имеет вид

$$U(x, t) = \sum_{i=1}^{N} \alpha_i (t) \varphi_i (x),$$

где базисные функции удовлетворяют граничному условию, т. е.

$$\varphi_i (0) = \varphi_i (l) = 0 \quad (i = 1, \ldots, N). \quad (6.21)$$

Тогда система уравнений (6.20) примет вид

$$\sum_{j=1}^{N} \left\{ \frac{d\alpha_j}{dt} d_{ij} + \alpha_j c_{ij} \right\} = 0 \quad (i = 1, \ldots, N), \quad (6.22)$$

где

$$c_{ij} = \int_0^l \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx}\, dx$$

и

$$d_{ij} = \int_0^l \varphi_i \varphi_j\, dx.$$

Хотя полудискретные методы разработаны теоретически главным образом в применении к параболическим уравнениям, они с таким же успехом могут применяться и к гиперболическим уравнениям, в особенности к уравнениям вида

$$\frac{\partial^2 u}{\partial t^2} + \lambda \frac{\partial u}{\partial t} + Au = 0 \quad (\lambda > 0),$$

которые описывают затухающие механические колебания. Для таких задач полудискретная аппроксимация вида

$$U(\mathbf{x}, t) = \sum_{i=1}^{N} \alpha_i(t)\, \varphi_i(\mathbf{x})$$

приводит к системе обыкновенных дифференциальных уравнений второго порядка относительно неизвестных функций $\alpha_i(t)$. Эту систему дополняют *две* последовательности начальных условий

$$\alpha_i(t_0) = c_i \quad (i = 1, \ldots, N)$$

и

$$\frac{d\alpha_i(t_0)}{di} = d_i \quad (i = 1, \ldots N);$$

коэффициенты c_i и d_i определяются заданными начальными значениями

$$u(x, t_0) = u_0(\mathbf{x})$$

и

$$\frac{\partial u(\mathbf{x}, t_0)}{\partial t} = v_0(\mathbf{x})$$

так, чтобы

$$\left\| u_0(\mathbf{x}) - \sum_{i=1}^{N} c_i \varphi_i(\mathbf{x}) \right\|^2_{\mathcal{L}_2(R)} = \min$$

и

$$\left\| v_0(\mathbf{x}) - \sum_{i=1}^{N} d_i \varphi_i(\mathbf{x}) \right\|^2_{\mathcal{L}_2(R)} = \min$$

соответственно,

Нелинейные задачи

В гл. 3 мы привели пример нелинейного уравнения $A(u)=f$, которое может быть решено методом Галеркина. В то же время нужно помнить, что преимущества метода Галеркина не могут быть полностью реализованы, если невозможно применить интегрирование по частям для упрощения скалярных произведений. Это верно и в том случае, когда мы применяем полудискретный метод Галеркина для решения нелинейного параболического уравнения

$$\frac{\partial u}{\partial t} + A(u) = 0.$$

Рассмотрим в качестве примера такой же нелинейный член, как и в эллиптическом случае, а именно

$$A(u) = -\frac{\partial}{\partial x}\left(p(u)\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(q(u)\frac{\partial u}{\partial y}\right).$$

Чтобы применить полудискретный метод Галеркина к этому уравнению, сначала необходимо переписать уравнение в слабой форме

$$\left(\frac{\partial U}{\partial t},\ \varphi_i\right) + a(U, \varphi_i) = 0 \quad (i = 1, \ldots, N).$$

Затем второй член упрощается с помощью интегрирования по частям, после чего получится система нелинейных обыкновенных дифференциальных уравнений вида

$$\sum_{j=1}^{N} \frac{d\alpha_j}{dt} d_{ij} + c_i(\alpha) = 0 \quad (i = 1, \ldots, N)$$

(ср. с (6.22) и с (6.37), где $\alpha = (\alpha_1, \ldots, \alpha_N)^T$. Вопрос о решении такой нелинейной системы обсуждается ниже в этой главе.

Упражнение 4. Опишите полудискретный метод Галеркина, построенный с помощью кусочно-линейных базисных функций, в применении к уравнению затухающих колебаний струны

$$\frac{\partial^2 u}{\partial t^2} + \lambda\frac{\partial u}{\partial t} - c^2\frac{\partial^2 u}{\partial x^2} = 0 \quad (0 < x < l;\ t_0 < t \leqslant t_1;\ \lambda > 0),$$

дополненному граничными и начальными условиями

$$u(0, t) = u(l, t) = 0 \quad (t_0 \leqslant t \leqslant t_1),$$

$$\left.\begin{aligned} u(x, t_0) &= u_0(x) \\ \frac{\partial u(x, t_0)}{\partial t} &= v_0(x) \end{aligned}\right\} (0 \leqslant x \leqslant l).$$

Упражнение 5. Опишите полудискретный метод в применении к уравнению

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \quad ((x, y, t) \in R \times (t_0, t_1]),$$

дополненному начальными и граничными условиями

$$\left.\begin{aligned} u(x, y, t_0) &= u_0(x, y) \\ \frac{\partial u(x, y, t_0)}{\partial t} &= v_0(x, y) \end{aligned}\right\} ((xy) \in R),$$

$$\frac{\partial u(x, y, t)}{\partial n} = f(x, y, t) \quad ((x, y, t) \in \partial R \times (t_0, t_1]),$$

где ∂R есть граница области $R = (x_0 < x < x_1) \times (y_0 < y < y_1)$, а n является внешней нормалью к границе.

Упражнение 6. Покажите, что применение полудискретного метода Галеркина к линейному дифференциальному уравнению

$$b_1 \frac{\partial^2 u}{\partial t^2} + c_1 \frac{\partial u}{\partial t} + Au = f$$

приводит к системе обыкновенных дифференциальных уравнений вида

$$P\ddot{\boldsymbol{\alpha}} + Q\dot{\boldsymbol{\alpha}} + R\boldsymbol{\alpha} = \mathbf{b},$$

где матрицы P, Q и R не зависят от времени и $c_1 P = b_1 Q$.

6.4. Непрерывные по времени методы

Как показано в предыдущем разделе, полудискретный метод приводит к системе обыкновенных дифференциальных уравнений

$$A\ddot{\boldsymbol{\alpha}} + B\dot{\boldsymbol{\alpha}} + C\boldsymbol{\alpha} = \mathbf{b} \tag{6.23}$$

с начальными условиями

$$\boldsymbol{\alpha}(t_0) = \mathbf{c}_0 \tag{6.24a}$$

и

$$\dot{\boldsymbol{\alpha}}(t_0) = \mathbf{d}_0, \tag{6.24b}$$

если $A \neq 0$. Для линейных задач, у которых матрицы A, B и C постоянны, решение уравнения (6.23) может быть получено стандартными аналитическими методами.

Если $A = 0$, то задача упрощается, и ее решение может быть записано в виде

$$\boldsymbol{\alpha}(t) = C^{-1}\mathbf{b} + \exp\left(-tB^{-1}C\right)\left(\mathbf{c}_0 - C^{-1}\mathbf{b}\right)$$

и выражено (Уэйт и Митчелл, 1971) в терминах решения задачи на собственные значения для уравнения

$$(\lambda B - C)\,\mathbf{u} = 0.$$

Если же $B = 0$, то $\boldsymbol{\alpha}(t)$ может быть выражено в терминах решения задачи на собственные значения для уравнения

$$(\lambda^2 A - C)\,\mathbf{u} = 0.$$

Хотя и можно получить полное решение отдельной задачи на собственные значения, для больших систем вычисления будут очень дорогостоящими, и поэтому в таких случаях часто выгоднее аппроксимировать решение уравнения (6.23) небольшим числом одних преобладающих компонент. Такие компоненты обычно очень слабо изменяются относительно изменений во времени и соответствуют наименьшим по модулю собственным значениям. Конкретные собственные значения вместе с соответствующими собственными векторами могут быть вычислены методом обратной итерации (Уилкинсон, 1965, стр. 534) значительно дешевле по сравнению с полным решением задачи на собственные значения, и поэтому такой подход обладает определенным преимуществом при условии, что аппроксимация немногими преобладающими компонентами адекватна решаемой задаче. Такая аппроксимация является особенно подходящей, если (I) $A \neq 0$ и необходимо *сглаживать* осцилляции или (II) $A = 0$ и требуется знать *стационарное* состояние, а не процесс его *установления*.

Рассмотрим для примера уравнение

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial y^2} \quad (0 < x,\ y < \pi;\quad t > 0), \tag{6.25}$$

решение которого удовлетворяет граничным условиям

$$u(x, 0, t) = u(x, \pi, t) = 0 \quad (0 \leqslant x \leqslant \pi), \tag{6.26a}$$

$$u(0, y, t) = 0 \quad (0 \leqslant y \leqslant \pi) \tag{6.26b}$$

и

$$u(\pi, y, t) = \sin y \quad (0 \leqslant y \leqslant \pi), \tag{6.26c}$$

а также начальному условию

$$u(x, y, 0) = \frac{x}{\pi} \sin y \quad (0 \leqslant x, y \leqslant \pi). \tag{6.26d}$$

Для построения конечноэлементного решения воспользуемся полудискретным методом и билинейным базисом. В данном случае необходимо специальным образом учесть то обстоятельство, что граничное условие (6.26c) является неоднородным (и гладким). Это можно сделать различными способами, и мы проведем некоторое сравнение их между собой. Будем

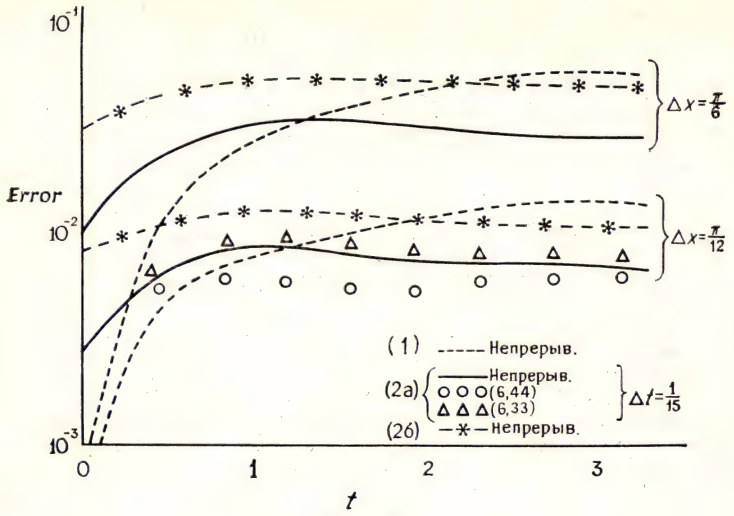


Рис. 27.

искать аппроксимирующее решение в виде

$$U(x, y, t) = V(x, y, t) + W(x, y), \tag{6.27}$$

где функция

$$V(x, y, t) = \sum_{i, j=1}^{N} V_{ij}(t)\,\varphi_{ij}(x, y) \tag{6.28}$$

есть кусочная билинейная аппроксимация решения уравнения

$$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} - \frac{x}{\pi}\sin y \quad (0 < x, y < \pi;\ t > 0) \tag{6.29}$$

с однородными начальными и граничными условиями, а $W(x, y)$ либо

(1) $$W(x, y) = \frac{x}{\pi}\sin y, \tag{6.30}$$

либо

(2) $$W(x, y) = \sum_{i, j=0}^{N+1} c_{ij}\varphi_{ij}(x, y), \tag{6.31}$$

и коэффициенты c_{ij} определяются с помощью начального условия.

Относительные достоинства различных способов вычисления такой аппроксимации видны из рис. 27. Точное решение задачи есть

$$u(x, y, t) = \frac{\operatorname{sh} x}{\operatorname{sh} \pi}\sin y + \frac{2}{\pi}\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k(1+k^2)^{1/2}}\, e^{-(1+k^2)^{1/2}} \sin kx \sin y.$$

Для способа (1) $W(x, y)$ задается равенством (6.30), для способа (2b) $W(x, y)$ задается в виде (6.31) и интерполирует начальное условие; для способа (2a) $W(x, y)$ задается в виде (6.31), причем коэффициенты c_{ij} выбираются так, чтобы получилась наилучшая аппроксимация начального условия в $\mathscr{L}_2(R)$.

6.5. Дискретизация по времени

Для нелинейных задач уже нельзя получить решения в таком виде, как в предыдущем разделе, и такими решениями редко пользуются для задач, в которых граничные условия зависят от времени. В таких случаях необходимо получать численное решение пошаговым методом. Подробное изложение численных методов для системы обыкновенных дифференциальных уравнений можно найти у многих авторов (см., например, Ламберт 1973), и мы рассмотрим только такие методы, которые являются подходящими для вычисления конечноэлементных решений. Система таких уравнений, как (6.23), может быть *жесткой* (Ламберт, 1973, стр. 231), а это означает, что они могут быть решены с удовлетворительной точностью только некоторыми специальными методами (Лаури, 1977, Гопкинс и Уэйт, 1976).

Параболические уравнения с частными производными и соответствующие им системы первого порядка (по времени), вероятно, заслуживают наибольшего внимания, и наиболее распространенным способом их решения является так называемый метод *Кранка — Николсона — Галеркина*. В этом методе система дифференциальных уравнений

$$B\dot{\boldsymbol{\alpha}} + C\boldsymbol{\alpha} = \mathbf{b} \tag{6.32}$$

заменяется системой разностных уравнений

$$B\left\{\frac{\boldsymbol{\alpha}_{n+1} - \boldsymbol{\alpha}_n}{\Delta t}\right\} + C\left\{\frac{\boldsymbol{\alpha}_{n+1} + \boldsymbol{\alpha}_n}{2}\right\} = \mathbf{b}\,(\tau_{n+1/2})\,(n = 0, 1, \ldots), \tag{6.33}$$

где $\boldsymbol{\alpha}_n$ аппроксимируют $\boldsymbol{\alpha}(t_0 + n\Delta t)$ и $\tau_{n+1/2} = t_0 + (n + {}^1/_2)\Delta t$ $(n = 0, 1, \ldots)$. Такую форму аппроксимации решения системы обыкновенных дифференциальных уравнений более точно можно назвать методом трапеций. Из (6.33) следует, что на каждом шаге вычислений необходимо решать систему линейных алгебраических уравнений для нахождения значений $\boldsymbol{\alpha}_{n+1}$. К сожалению, ситуация не так проста для нелинейных уравнений вида

$$B\boldsymbol{\alpha} + \mathbf{a}(\boldsymbol{\alpha}) = \mathbf{b}, \tag{6.34}$$

которые получаются в результате применения полудискретного метода Галеркина к уравнениям вида

$$\frac{\partial u}{\partial t} + A(u) = f. \tag{6.35}$$

Аппроксимация Кранка — Николсона — Галеркина в случае (6.34) сведется к системе нелинейных уравнений для определения α_{n+1}, так что придется применять метод «*предиктор — корректор*». Дуглас и Дюпон (1970) предложили ряд различных схем и провели их сравнительный анализ, но здесь мы можем лишь вкратце остановиться на типичном примере предлагаемых ими методов для решения уравнения (6.35) при

$$A(u) = -\frac{\partial}{\partial x}\left(p(u)\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(q(u)\frac{\partial u}{\partial y}\right). \tag{6.36}$$

В этом случае $\mathbf{c}(\boldsymbol{\alpha}) = (c_1(\boldsymbol{\alpha}), \ldots, c_N(\boldsymbol{\alpha}))^{\mathrm{T}}$ и

$$c_i(\boldsymbol{\alpha}) = \sum_{j=1}^{N} \alpha_j \iint_R \left\{ p(U)\frac{\partial \varphi_i}{\partial x}\frac{\partial \varphi_j}{\partial x} + q(U)\frac{\partial \varphi_i}{\partial y}\frac{\partial \varphi_j}{\partial y} \right\} dx\,dy \tag{6.37}$$

$$(i = 1, \ldots, N).$$

Поэтому $\mathbf{c}(\boldsymbol{\alpha})$ можно представить как

$$\mathbf{c}(\boldsymbol{\alpha}) = D(\boldsymbol{\alpha})\boldsymbol{\alpha}$$

и вместо (6.34) получить

$$B\dot{\boldsymbol{\alpha}} + D(\boldsymbol{\alpha})\boldsymbol{\alpha} = \mathbf{b};$$

тогда (6.33) заменится двумя уравнениями

$$B\left\{\frac{\boldsymbol{\beta}_{n+1} - \boldsymbol{\alpha}_n}{\Delta t}\right\} + D(\boldsymbol{\alpha}_n)\left\{\frac{\boldsymbol{\beta}_{n+1} + \boldsymbol{\alpha}_n}{2}\right\} = \mathbf{b}(\tau_{n+1/2}) \tag{6.38}$$

и

$$B\left\{\frac{\boldsymbol{\alpha}_{n+1} - \boldsymbol{\alpha}_n}{\Delta t}\right\} + D\left(\frac{\boldsymbol{\beta}_{n+1} + \boldsymbol{\alpha}_n}{2}\right)\left\{\frac{\boldsymbol{\alpha}_{n+1} + \boldsymbol{\alpha}_n}{2}\right\} = \mathbf{b}(\tau_{n+1/2}). \tag{6.39}$$

Предиктор (6.38) дает первое приближение $\boldsymbol{\beta}_{n+1}$, а затем можно дополнительно использовать корректор (6.39) для улучшения этого приближения.

Упражнение 7. Покажите, что полудискретный метод, использующий аппроксимацию Кранка — Николсона — Галеркина, в применении к простейшему уравнению диффузии

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \quad (0 < x < l;\ t_0 < t \leqslant t_1)$$

в случае линейных базисных функций сводится к системе разностных уравнений

$$\Delta_t I_x U_i^n - \frac{r}{2}\,\delta_x^2\{U_i^{n+1} + U_i^n\} = 0, \tag{6.40}$$

где операторы Δ_t, I_x и δ_x^2, постоянная r и разбиение области такие же, как и в упр. 3 из разд. 6.2.

Одношаговый метод Галеркина (*конечные элементы по времени*)

Другой подход состоит в дискретизации уравнения (6.32) с помощью аппроксимации Галеркина, т. е. в получении приближенного решения в виде

$$\boldsymbol{\alpha}^{(n)}(t) = \sum_{j=0}^{S} \boldsymbol{\alpha}_j^{(n)} \varphi_j^{(n)}(t) \quad (n = 0,\ 1,\ \ldots) \tag{6.41}$$

на каждом подынтервале $(\tau_n,\ \tau_n + \Delta t)$, где коэффициенты $\boldsymbol{\alpha}_j^{(n)}$ $(j = 1, \ldots, S)$ определяются из системы

$$\langle B\dot{\boldsymbol{\alpha}}^{(n)} + C\boldsymbol{\alpha}^{(n)},\ \varphi_j^{(n)}\rangle_n = \langle \mathbf{b},\ \varphi_j^{(n)}\rangle_n, \quad (j = 1, \ldots, S;\ n = 0,\ 1, \ldots) \tag{6.42}$$

и условия непрерывности

$$\boldsymbol{\alpha}^{(n)}(\tau_n^+) = \boldsymbol{\alpha}^{(n-1)}(\tau_n^-) \quad (n = 1,\ 2,\ \ldots). \tag{6.43}$$

i-я компонента *вектора* $\langle \mathbf{u}, v\rangle_n$ есть

$$\int_{\tau_n}^{\tau_n + \Delta\tau} u_i(t)\, v(t)\, dt,$$

где $\mathbf{u} = (u_1(t), \ldots, u_N(t))^{\mathrm{T}}$.

Заметим, что в представление (6.41) входят $S + 1$ базисных функций, тогда как система (6.42) содержит только S уравнений, и поэтому вид разностной аппроксимации зависит от способа упорядочения базисных функций.

Разобьем подынтервал $[\tau_n, \tau_n + \Delta t]$ так, чтобы

$$\tau_n = \tau_0^{(n)} < \tau_1^{(n)} < \ldots < \tau_S^{(n)} = \tau_n + \Delta t$$

и

$$\tau_j^{(n)} - \tau_{j-1}^{(n)} = \frac{\Delta t}{S} \quad (j = 1, \ldots, S).$$

Предположим, что $\varphi_j^{(n)}$ $(j=0, 1, \ldots, S)$ образуют базис для лагранжевой интерполяции на $[\tau_n, \tau_n + \Delta t]$, т. е.

$$\varphi_j^{(n)}(\tau_k^{(n)}) = \begin{cases} 1 & (j=k) \\ 0 & (j \neq k) \end{cases} \quad (n=0, 1, \ldots),$$

и $\varphi_j^{(n)}(t)$ есть полиномы степени S на $[\tau_n, \tau_n + \Delta t]$. Из (6.43) следует, что

$$\boldsymbol{\alpha}_0^{(n)} = \boldsymbol{\alpha}_S^{(n-1)} = \boldsymbol{\alpha}_n \approx \boldsymbol{\alpha}(\tau_n) \quad (n=1, 2, \ldots),$$

и поэтому при $S \geqslant 2$ можно исключить $\boldsymbol{\alpha}_j^{(n)}$ $(j=1, \ldots, S-1)$ из (6.42) и получить одно уравнение, связывающее $\boldsymbol{\alpha}_{n+1}$ и $\boldsymbol{\alpha}_n$ $(n=0, 1, \ldots)$.

Рассмотрим, например, тот случай, когда $\mathbf{b}=0$:

(1) $S=1$ (Комини, дель Гвидичи, Левис и Зенкевич, 1974). Тогда

$$\left\{B + \tfrac{2}{3}\Delta t C\right\}\boldsymbol{\alpha}_{n+1} = \left\{B - \tfrac{\Delta t}{3} C\right\}\boldsymbol{\alpha}_n. \tag{6.44}$$

(2) $S=2$. Тогда

$$\left\{I + \tfrac{3}{5}\Delta t M + \tfrac{3}{20}(\Delta t)^2 M^2\right\}\boldsymbol{\alpha}_{n+1} = \left\{I - \tfrac{2}{5}\Delta t M + \tfrac{1}{20}(\Delta t)^2 M^2\right\}\boldsymbol{\alpha}_n,$$

где $M = B^{-1}C$.

Можно показать (Халм, 1972), что некоторые хорошо известные разностные методы могут быть сформулированы как одношаговые методы Галеркина.

При использовании эрмитовой интерполяции получатся другие формулы. Другим возможным способом получения разностных схем является замена аппроксимации Галеркина (6.42) аппроксимацией в смысле метода наименьших квадратов.

Упражнение 8. Покажите, что если базисные функции расположены в обратном порядке, т. е.

$$\varphi_j^{(n)}(\tau_{S-k}^{(n)}) = \begin{cases} 1 & (k=j) \\ 0 & (k \neq j) \end{cases} \quad (n=0, 1, \ldots),$$

и если $\mathbf{b}=0$ и $S=1$, то разностное уравнение примет вид

$$\left\{B + \tfrac{\Delta t}{3} C\right\}\boldsymbol{\alpha}_{n+1} = \left\{B - \tfrac{2}{3}\Delta t C\right\}\boldsymbol{\alpha}_n.$$

Упражнение 9. Можно получить другие системы разностных уравнений, если определить локальную аппроксимацию (6.41) с помощью системы

$$\langle B\dot{\boldsymbol{\alpha}}^{(n)} + C\boldsymbol{\alpha}^{(n)}, \psi_j^{(n)}\rangle_n = \langle \mathbf{b}, \psi_j^{(n)}\rangle_n \quad (j=1, \ldots, S;\ n=0, 1, \ldots)$$

вместо (6.42), где $\{\varphi_j^{(n)}\}$ и $\{\psi_j^{(n)}\}$ отличны друг от друга. Покажите, что при $\mathbf{b} = 0$,

$$\psi_1^{(n)}(t) = 1 \quad (n = 0, 1, \ldots;\ t \geqslant t_0)$$

и

$$\psi_2^{(n)}(t) = \frac{2}{\Delta t}(t - \tau_n) - 1 \quad (n = 0, 1, \ldots;\ \tau_{n-1} \leqslant t \leqslant \tau_n)$$

получатся следующие разностные уравнения:

(1) $S = 1$. Тогда

$$\left\{B + \frac{\Delta t}{2} C\right\} \boldsymbol{\alpha}_{n+1} = \left\{B - \frac{\Delta t}{2} C\right\} \boldsymbol{\alpha}_n.$$

(2) $S = 2$. Тогда

$$\left\{I + \frac{\Delta t}{2} M + \frac{(\Delta t)^2}{12} M^2\right\} \boldsymbol{\alpha}_{n+1} = \left\{I - \frac{\Delta t}{2} M + \frac{(\Delta t)^2}{12} M^2\right\} \boldsymbol{\alpha}_n,$$

где $M = B^{-1}C$.

Упражнение 10. Покажите, что если заменить (6.42) аппроксимацией в смысле наименьших квадратов, то при $S = 1$ и $\mathbf{b} = 0$ получится разностное уравнение

$$\left\{\frac{1}{\Delta t} B^2 + \frac{1}{2}[BC + CB] + \frac{\Delta t}{3} C^2\right\} \boldsymbol{\alpha}_{n+1} =$$

$$= \left\{\frac{1}{\Delta t} B^2 + \frac{1}{2}[BC - CB] - \frac{\Delta t}{6} C^2\right\} \boldsymbol{\alpha}_n.$$

Метод переменных направлений Галеркина для параболических уравнений (Денди и Файервезер, 1975).

В гл. 3 этот метод был применен к эллиптическим задачам. Аналогичным образом его можно применять и к параболическим задачам, определенным на прямоугольных областях. Если используемые базисные функции заданы в виде тензорного произведения, то для аппроксимации, определяемой линейным уравнением

$$\left(\frac{\partial U}{\partial t},\ V\right) + a(U,\ V) = (f,\ U) \quad (t > t_0),$$

получается (в случае двумерной задачи) алгебраическая система, которая может быть записана как (см. гл. 3, стр. 71)

$$B_x \otimes B_y \left\{\frac{\boldsymbol{\alpha}_{n+1} - \boldsymbol{\alpha}_n}{\Delta t}\right\} + (A_x \otimes B_y + B_x \otimes A_y) \left\{\frac{\boldsymbol{\alpha}_{n+1} + \boldsymbol{\alpha}_n}{2}\right\} = \mathbf{b}_{n+1/2}$$

$$(n = 1, 2, \ldots), \qquad (6.45)$$

где через $\otimes$ обозначается тензорное произведение. Если к левой части добавить член $\left(\frac{1}{2}\Delta t\right)^2 A_x \otimes A_y \boldsymbol{\alpha}_{n+1}$, то (6.45) можно

заменить уравнением

$$\left(B_x + \frac{\Delta t}{2} A_x\right) \otimes \left(B_y + \frac{\Delta t}{2} A_y\right) \alpha_{n+1} = \psi,$$

которое может быть решено в два этапа, как и в эллиптическом случае (разд. 3.4).

Был предложен и ряд других методов, но при этом во многих случаях рассматривались только линейные задачи, например, в приложениях общих многошаговых методов (Зламал, 1975) и методов Норсетта (Семенич и Глэдвелл, 1974). Дюпон, Файервезер и Джонсон (1974) построили семейства трехслойных разностных схем для решения как линейных, так и нелинейных задач. Файервезер и Джонсон (1975) показали, что можно использовать локальную экстраполяцию Ричардсона, основанную на таких трехслойных схемах, а также на некоторых двухслойных схемах, предложенных Дугласом и Дюпоном (1970). Они рассмотрели также влияние интерполяции нелинейных коэффициентов, т. е. интерполяции таких функций, как $p(u)$ и $q(u)$ в (6.36).

6.6. Сходимость полудискретных аппроксимаций Галеркина

Этот раздел содержит краткое изложение одной из оценок сходимости, полученных Томе и Уолбином (1975) и Уилером (1973) для модельной задачи, определяемой уравнением

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \quad ((x, y, t) \in R \times (t_0, t_1]),$$

начальным условием

$$u(x, y, t_0) = u_0(x, y) \quad ((x, y) \in R)$$

и граничным условием

$$u(x, y, t) = 0 \quad ((x, y, t) \in \partial R \times (t_0, t_1]).$$

При этом ограничение двумя измерениями не является существенным.

Отправная точка для анализа та же, что и в гл. 5; а именно, формулировка предположения об аппроксимирующих свойствах подпространства K_N. Поэтому предположим, что теорема 5.4 справедлива и существует такое $k \geqslant 1$, что для любого $u \in \mathscr{H}_2^{(k+1)}(R)$ ошибка, порождаемая интерполяцией u элементом $\tilde{u} \in K_N$, ограничена как

$$\| u - \tilde{u} \|_{r, R} \leqslant C h^{k+1-r} \| u \|_{k+1, R} \quad (r \leqslant k). \tag{6.46}$$

Если предположить, что нарушения вариационных принципов, отмеченные в гл. 5, отсутствуют, то полудискретная

аппроксимация Галеркина будет удовлетворять уравнению

$$\left(\frac{\partial U}{\partial t},\ V\right)+a(U,\ V)=0 \quad (\text{для всех } V\in \mathring{K}_N). \tag{6.47}$$

Если $u\in \mathring{\mathscr{H}}_2^{(k+1)}(R)\times C^1[t_0,\ t_1]$, то из (6.46) следует, что *проекция* $W\in \mathring{K}_N\times C^1[t_0,\ t_1]$, которая при $t\in[t_0,\ t_1]$ удовлетворяет условию

$$a(W,\ V)=a(u,\ V) \quad (\text{для всех } V\in \mathring{K}_N), \tag{6.48}$$

будет также удовлетворять неравенству

$$\|u-W\|_{r,\ R}\leqslant Ch^{k+1-r}\|u\|_{k+1,\ R}\ (t\in[t_0,\ t_1]). \tag{6.49}$$

Из (6.47) и (6.48) получим, что

$$\left(\frac{\partial U}{\partial t}-\frac{\partial W}{\partial t},\ V\right)+a(U-W,\ V)=\left(\frac{\partial u}{\partial t}-\frac{\partial W}{\partial t},\ V\right)$$

для любого $t\in[t_0,\ t_1]$; при $V=U_t-W_t$ это дает

$$\|U_t-W_t\|^2_{\mathscr{L}_2(R)}+\frac{1}{2}\frac{d}{dt}a(U-W,\ U-W)=(u_t-W_t,\ U_t-W_t).$$

Применяя неравенство Шварца к правой части, получим

$$\|U_t-W_t\|^2_{\mathscr{L}_2(R)}+\frac{1}{2}\frac{d}{dt}a(U-W,\ U-W)\leqslant$$
$$\leqslant\|u_t-W_t\|_{\mathscr{L}_2(R)}\|U_t-W_t\|_{\mathscr{L}_2(R)}.$$

Из (6.49) следует, что

$$\|u_t-W_t\|_{\mathscr{L}_2(R)}\leqslant Ch^k\|u\|_{k+1,\ R},$$

так что

$$\frac{1}{2}\frac{d}{dt}a(U-W,\ U-W)\leqslant Ch^{2k}\|u\|^2_{k+1,\ R},$$

и поэтому

$$a(U-W,\ U-W)\leqslant Ch^{2k},\quad C\equiv C(u). \tag{6.50}$$

Так как билинейная форма a эллиптична в $\mathring{\mathscr{H}}_2^{(1)}$ (разд. 5.2), то объединение (6.49) и (6.50) дает

$$\|u-U\|_{1,\ R}\leqslant Ch^k,\quad C\equiv C(u).$$

Это оценка в непрерывной форме для аппроксимации Галеркина. Если уравнение (6.47) решается пошаговым методом, то должен быть рассмотрен дополнительный источник ошибок. Рядом авторов были получены оценки таких ошибок; дальнейшие ссылки по этому вопросу можно найти, например, у Денди (1975) или у де Бура (1974).

ГЛАВА 7

ДАЛЬНЕЙШЕЕ РАЗВИТИЕ ТЕОРИИ И ПРИЛОЖЕНИЯ

7.1. Введение

Прежде чем применять метод конечных элементов в различных его формах к решению задач, полезно в сжатой форме сформулировать основные характеристики этого метода.

Метод может быть использован для решения как стационарных, так и нестационарных задач. Ограниченная пространственная или пространственно-временная область разбивается на некоторое число неперекрывающихся элементов. Аппроксимирующие функции, которые могут быть полиномами, рациональными дробями и т. д., относятся к конкретным элементам, и параметры при этих аппроксимирующих функциях согласованы так, чтобы обеспечивалась желаемая степень гладкости аппроксимации на стыках между соседними элементами. Тогда аппроксимирующая функция во всей области может быть выражена с помощью своих значений и значений своих производных в узловых точках области через базисные функции, которые отличны от нуля только на немногих элементах, расположенных вокруг соответствующих узлов. Более точно, аппроксимирующая функция для всей области имеет вид

$$U(\mathbf{x}) = \sum_{i=1}^{N} \left[p_i(\mathbf{x})\, U_i + q_i(\mathbf{x}) \left(\frac{\partial U}{\partial x_1} \right)_i + r_i(\mathbf{x}) \left(\frac{\partial U}{\partial x_2} \right)_i + \ldots \right], \quad (7.1)$$

где $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, функции $p_i(\mathbf{x})$, $q_i(\mathbf{x})$, $r_i(\mathbf{x})$ и т. д. имеют локальный носитель, а N есть число узловых точек в области. Функции $p_i(\mathbf{x})$, $\partial q_i(\mathbf{x})/\partial x_1$, $\partial r_i(\mathbf{x})/\partial x_2$ и т. д. принимают единичное значение в узле i. Во многих случаях (7.1) имеет упрощенный вид

$$U(\mathbf{x}) = \sum_{i=1}^{N} p_i(\mathbf{x})\, U_i, \quad (7.2)$$

хотя есть задачи, для которых необходимо использовать представление (7.1) общего вида, особенно те, в которых требуется бóльшая гладкость между элементами или повышенная точность в определении градиента решения. Построение базисных функций $p_i(\mathbf{x})$, $q_i(\mathbf{x})$, $r_i(\mathbf{x})$ и т. д. является одним из наиболее важных и часто одним из самых трудных моментов

в методе конечных элементов. Это в особенности верно для задач с криволинейными границами и линиями раздела, особенностями и т. д., и для задач с производными высоких порядков. Вопросы построения базисных функций изложены в гл. 4.

В конце этой главы мы увидим, как решаются некоторые, в основном физические и инженерные задачи методом конечных элементов в различных его формах (Ритца, Галеркина, наименьших квадратов, колокации). Разнообразные типы базисных функций и модификации основного метода будут использованы для того, чтобы подчеркнуть относительные преимущества тех различных приемов, которые объединяются под общим названием метода конечных элементов. Что касается базисных функций для задач, в которых требуется высокая степень гладкости между элементами (например, решение бигармонического уравнения в смысле наименьших квадратов должно принадлежать пространству C^3), то здесь будут применены несогласованные элементы, и поэтому мы начнем эту главу с краткого описания некоторых используемых на практике несогласованных элементов.

7.2. Несогласованные элементы [1])

До сих пор аппроксимация для всей области в методе конечных элементов строилась в предположении ее некоторой гладкости (или по крайней мере непрерывности) на стыках между соседними элементами. Для дифференциального уравнения порядка $2k$ требовалась сшивка в C^{k-1} для методов Ритца и Галеркина или сшивка в C^{2k-1} для метода наименьших квадратов. Если для тетраэдральных элементов сшивка в C^1 достигается применением полиномов девятой степени, то нетрудно себе представить, каким сложным делом будет при $k > 1$ построение элементов с требуемой степенью гладкости сшивки, т. е. построение согласованных элементов. Поэтому с вычислительной точки зрения желательно научиться использовать элементы с меньшей степенью гладкости на стыках, чем это формально требуется, т. е. несогласованные элементы.

Поскольку инженеры в меньшей степени по сравнению с математиками избегают применения на практике теоретически не обоснованных приемов, нет ничего удивительного в том, что несогласованные элементы были впервые предложены именно инженерами. Ими было предложено также так называемое кусочное тестирование для выбора таких несогласован-

---

[1]) Авторы признательны проф. Р. Барнхиллу и Дж. Брауну за полезные дискуссии при подготовке материала этого раздела.

ных элементов, которые обеспечивали бы сходимость конечноэлементной аппроксимации для данной задачи. В действительности кусочное тестирование является проверкой непротиворечивости несогласованного конечноэлементного метода, используемого для решения конкретной задачи.

Кусочное тестирование (Айронс и Раззак, 1972)

Смысл кусочного тестирования состоит в следующем. Предположим, что пространство несогласованных базисных функций содержит все полиномы такого порядка r, какой имеет старшая производная в энергетическом функционале (в обозначениях гл. 5 $P_r \subset K_h$), и пусть граничные условия вдоль периметра произвольно взятой части элементов определены как значения произвольно взятого частного решения $u \in P_r$ на этой линии. Тогда кусочное тестирование считается выполненным, если приближенное решение U_h, вычисленное по методу конечных элементов в форме Ритца без учета разрывов на границах между элементами, совпадает с u на рассматриваемой части элементов. Таким образом, кусочное тестирование выполнено, если при $u \in P_r$

$$U_h = u. \tag{7.3}$$

В качестве иллюстрации применения кусочного тестирования к задачам второго порядка рассмотрим решение уравнения Лапласа на части элементов, образующей единичный квадрат и составленной из двух треугольных элементов (рис. 28). Энергетический функционал содержит первые производные, и поэтому $r = 1$. На каждом треугольнике определим линейную функцию по ее значениям в серединах трех его

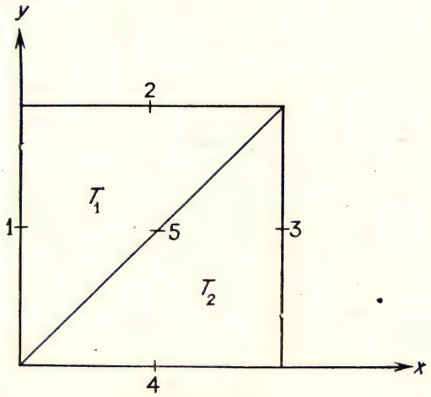


Рис. 28.

сторон. Тогда

$$U^{[1]}(x, y) = (1-2x)U_1 - (1-2y)U_2 + (1+2x-2y)U_5$$

и

$$U^{[2]}(x, y) = -(1-2x)U_3 + (1-2y)U_4 + (1-2x+2y)U_5$$

для треугольников T_1 и T_2 соответственно. Интерполянт на всем квадрате в общем случае разрывен вдоль границы раздела треугольников, и поэтому такие элементы являются несогласованными для метода Ритца.

В качестве примера возьмем на этой части элементов тестовое решение

$$u = x + y,$$

принимающее граничные значения

$$u_1 = u_4 = \frac{1}{2}, \quad u_2 = u_3 = \frac{3}{2},$$

так что указанные интерполянты запишутся в виде

$$U^{[1]}(x, y) = (-1-x+3y) + (1+2x-2y)U_5 \qquad (7.4a)$$

и

$$U^{[2]}(x, y) = (-1+3x-y) + (1-2x+2y)U_5. \qquad (7.4b)$$

Метод конечных элементов в форме Ритца сводится к минимизации энергетического функционала

$$\iint_{T_1} (U_x^{[1]2} + U_y^{[1]2})\,dx\,dy + \iint_{T_2} (U_x^{[2]2} + U_y^{[2]2})\,dx\,dy$$

относительно U_5, что дает

$$U_5 = 1.$$

Подставляя это значение в (7.4a) и в (7.4b), получим

$$U^{[1]}(x, y) = U^{[2]}(x, y) = x + y,$$

и поэтому

$$U_h = u$$

на всей рассматриваемой части. В действительности же это верно для любого $u \in P_1$ и любой части таких элементов, т. е. такой несогласованный элемент выдерживает кусочное тестирование.

Упражнение 1. Проведите кусочное тестирование для элементов, изображенных на рис. 29, и покажите, что

$$U_5 = \frac{1-\alpha-\alpha^2+2\alpha^3}{1-2\alpha+2\alpha^3}.$$

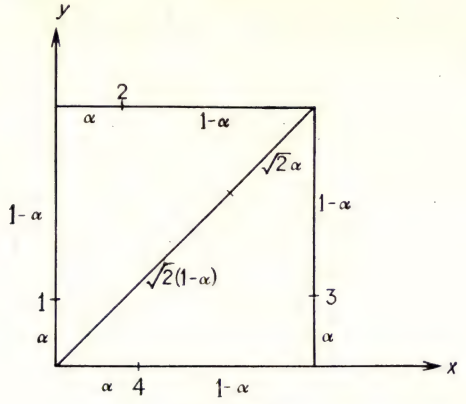


Рис. 29.

Докажите тем самым, что такие элементы выдерживают кусочное тестирование только при $\alpha = {}^1/_2$.

В качестве следующего примера применения кусочного тестирования рассмотрим задачу четвертого порядка, определяемую на квадратной области бигармоническим уравнением и заданием решения и его нормальной производной на границе. Разобьем квадрат обычным образом на прямоугольные треугольники равной площади и снова рассмотрим часть элементов в виде единичного квадрата, состоящего из двух треугольных элементов (рис. 30). Для бигармонического уравнения энергетический функционал содержит вторые производные, и поэтому $r = 2$. На каждом треугольнике определим квадратичную функцию по ее значениям в вершинах треугольника и по значениям ее нормальной производной в серединах

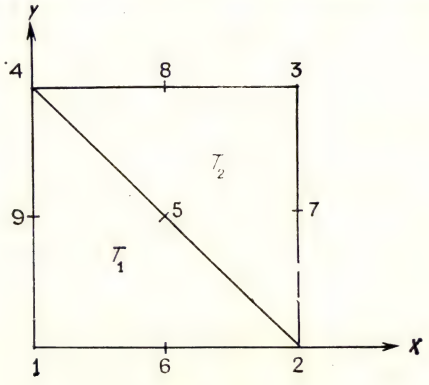


Рис. 30.

сторон. Такой элемент называется *треугольником Морли*. Интерполянты будут иметь вид

$$U^{[1]}(x, y) = (1 - x - y + 2xy)\,U_1 + \frac{1}{2}(x + y + x^2 - 2xy - y^2)\,U_2 + \\ + \frac{1}{2}(x + y - x^2 - 2xy + y^2)\,U_4 + y(1 - y)\left(\frac{\partial U}{\partial y}\right)_6 - \\ - \frac{1}{\sqrt{2}}(x + y - x^2 - 2xy - y^2)\left(\frac{\partial U}{\partial n}\right)_5 + x(1 - x)\left(\frac{\partial U}{\partial x}\right)_9$$

и

$$U^{[2]}(x, y) = \frac{1}{2}(3x - y + y^2 - x^2 - 2xy)\,U_2 + (1 - x - y + 2xy)\,U_3 + \\ + \frac{1}{2}(3y - x - y^2 + x^2 - 2xy)\,U_4 + x(x - 1)\left(\frac{\partial U}{\partial x}\right)_7 + \\ + y(y - 1)\left(\frac{\partial U}{\partial y}\right)_8 + \\ + \frac{1}{\sqrt{2}}(-2 + 3x + 3y - x^2 - 2xy - y^2)\left(\frac{\partial U}{\partial n}\right)_5$$

для треугольников T_1 и T_2 соответственно, где n есть внешняя нормаль по отношению к T_1 и внутренняя нормаль по отношению к T_2. Снова интерполянт на всем единичном квадрате в общем случае разрывен вдоль границы между двумя треугольниками, и поэтому такие элементы являются несогласованными. Для задачи четвертого порядка элементы останутся несогласованными и тогда, когда интерполянт будет непрерывным на этой границе, а его нормальная производная к ней — нет.

В качестве примера возьмем на этой части элементов тестовое решение

$$u = x^2 + y^2,$$

принимающее граничные значения

$$u_1 = 0, \quad u_2 = u_4 = 1, \quad u_3 = 2;$$

$$\left(\frac{\partial u}{\partial y}\right)_6 = \left(\frac{\partial u}{\partial x}\right)_9 = 0, \quad \left(\frac{\partial u}{\partial x}\right)_7 = \left(\frac{\partial u}{\partial y}\right)_8 = 2;$$

тогда указанные интерполянты запишутся как

$$U^{[1]}(x, y) = (x + y - 2xy) - \frac{1}{\sqrt{2}}(x + y)(1 - x - y)\left(\frac{\partial u}{\partial n}\right)_5 \quad (7.5a)$$

и

$$U^{[2]}(x, y) = (2 - 3x - 3y + 2x^2 + 2xy + 2y^2) + \\ + \frac{1}{\sqrt{2}}(x + y - 2)(1 - x - y)\left(\frac{\partial u}{\partial n}\right)_5. \quad (7.5b)$$

Метод конечных элементов в форме Ритца сводится к минимизации энергетического функционала

$$\iint_{T_1} (U_{xx}^{[1]^2} + 2U_{xy}^{[1]^2} + U_{yy}^{[1]^2})\,dx\,dy + \iint_{T_2} (U_{xx}^{[2]^2} + 2U_{xy}^{[2]^2} + U_{yy}^{[2]^2})\,dx\,dy$$

относительно параметра $(\partial u/\partial n)_5$, что дает

$$\left(\frac{\partial u}{\partial n}\right)_5 = \sqrt{2}.$$

Подставляя это значение в (7.5a) и в (7.5b), получим

$$U^{[1]}(x,\ y) = U^{[2]}(x,\ y) = x^2 + y^2,$$

и поэтому

$$U_h = u$$

на всей рассматриваемой части. В действительности же это верно для любого $u \in P_2$ и для любой части таких элементов, т. е. треугольник Морли выдерживает кусочное тестирование.

Упражнение 2. Покажите, что треугольный элемент, на котором полная квадратичная функция определяется своими значениями в вершинах и в серединах сторон, не выдерживает кусочного тестирования для задачи четвертого порядка, описанной выше.

Хотя математическая проверка несогласованных элементов кусочным тестированием привлекательна сама по себе, с практической точки зрения достаточно осуществить такую проверку на вычислительной машине. Элементы считаются выдержавшими кусочное тестирование, если численное решение воспроизводит заранее известный ответ с учетом, конечно, влияния ошибок округления.

Необходимое и достаточное условие сходимости построенной на несогласованных элементах аппроксимации, которое эквивалентно кусочному тестированию, в обозначениях разд. 5.4(E) имеет вид равенства

$$a_h(p,\ V_h) = (f,\ V_h) \quad \text{(при всех } V_h \in K_h) \tag{7.6}$$

для любого полиномиального решения $p \in P_r$, где $P_r \subset K_h$, а r есть порядок старшей производной, входящей в a_h.

Но любое решение u удовлетворяет уравнению

$$a(u,\ v) = (f,\ v)$$

при всех допустимых функциях $v \in \mathscr{H}$, и если оно также удовлетворяет уравнению

$$a(u,\ V_h) = (f,\ V_h) \tag{7.7}$$

при всех несогласованных функциях $V_h \in K_h$, то (7.6) примет вид

$$a_h(p, V_h) = a(p, V_h) \quad (\text{при всех } V_h \in K_h) \tag{7.8}$$

(Стренг и Фикс, 1973, с. 207). Например, если задача сводится к минимизации

$$I(v) = \iint_R \left[\left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right] dx\, dy,$$

то

$$a(p, V_h) = \iint_R \left\{\left(\frac{\partial p}{\partial x}\right)\left(\frac{\partial V_h}{\partial x}\right) + \left(\frac{\partial p}{\partial y}\right)\left(\frac{\partial V_h}{\partial y}\right)\right\} dx\, dy.$$

Если область R разбита на неперекрывающиеся элементы T_j $(j = 1, 2, \ldots, S)$, то тогда

$$a_h(p, V_h) = \sum_{j=1}^{S} \iint_{T_j} \left\{\left(\frac{\partial p}{\partial x}\right)\left(\frac{\partial V_h}{\partial x}\right) + \left(\frac{\partial p}{\partial y}\right)\left(\frac{\partial V_h}{\partial y}\right)\right\} dx\, dy.$$

Физический смысл равенства (7.8) состоит в том, что разрывы на границах между элементами можно не учитывать при вычислении $a(p, V_h)$.

Для установления эквивалентности (7.3) и (7.8) удобно ввести полунорму

$$|u|_h = [a_h(u, u)]^{1/2}.$$

Тогда можно получить оценки

$$|u - U_h|_h \geqslant \sup_{V_h} \left\{\frac{|a_h(u, V_h) - a(u, V_h)|}{|V_h|_h}\right\} \tag{7.9}$$

и (ср. с (5.16))

$$|u - U_h|_h \leqslant \inf_{V_h} |u - V_h|_h + \sup_{V_h} \frac{|a_h(u, V_h) - a(u, V_h)|}{|V_h|_h} \tag{7.10}$$

при условии, что выполнено (7.7). Тогда (7.8) следует из (7.3) и (7.9). Обратно, из (7.8) и (7.10) следует, что

$$|p - U_h|_h \leqslant \inf_{V_h} |p - V_h|_h.$$

Правая часть этого неравенства равна нулю, так как $P_r \subset \subset K_h$, и мы получаем (7.3).

Упражнение 3. Убедитесь в справедливости (7.7) для несогласованных кусочно-линейных элементов с узлами в серединах сторон.

Для задач второго порядка можно дать следующую полезную для практики формулировку кусочного тестирования: кусочное тестирование выдержано, если

$$\int_E (V_h^{[1]} - V_h^{[2]})\, d\sigma = 0, \tag{7.11}$$

где E есть любое внутреннее прямое ребро сетки, а V_h есть любая несогласованная функция, так что $V_h^{[1]}$ и $V_h^{[2]}$ являются ее предельными значениями по разные стороны ребра E (Браун, 1975).

В заключение отметим два несогласованных прямоугольных элемента, которые выдерживают кусочное тестирование:

(1) *Элемент Вильсона* (Вильсон и др., 1971). На квадрате $0 \leqslant x,\ y \leqslant 1$ шесть функций образуют базис: четыре билинейные xy, $x(1-y)$, $y(1-x)$, $(1-x)(1-y)$ и две дополнительные $4x(1-x)$ и $4y(1-y)$. Последние функции дают возможность представить в этом базисе произвольный квадратичный полином от двух переменных и тем самым повысить точность аппроксимации на каждом элементе.

(2) *Элемент Адини* (Адини и Клаф, 1961). Для этого элемента с 12 степенями свободы неизвестными параметрами являются значения u, $\partial u/\partial x$ и $\partial u/\partial y$ в вершинах квадрата, а входящие в полный кубический полином функции вместе с x^3y и xy^3 образуют базис.

Упражнение 4. Покажите, что прямоугольные элементы Вильсона и Адини выдерживают кусочное тестирование для задач второго и четвертого порядка соответственно.

7.3. Смешанные интерполянты

Один из методов получения конечноэлементных аппроксимаций, *точно* удовлетворяющих граничным условиям Дирихле, состоит в том, что в решение задачи включается некоторый смешанный функциональный интерполянт, построенный по заданным граничным значениям (Гордон, 1971). В простейшем случае — это билинейный смешанный интерполянт на квадрате.

Если, например, $f \in C^{2,2}(\bar{R})$, где $\bar{R} = [0, h] \times [0, h]$, то функция

$$\tilde{f}(x, y) = \left(1 - \frac{x}{h}\right) f(0, y) + \frac{x}{h} f(h, y) + \left(1 - \frac{y}{h}\right) f(x, 0) + \\ + \frac{y}{h} f(x, h) - \tilde{\tilde{f}}(x, y), \tag{7.12}$$

где

$$\tilde{\tilde{f}}(x, y) = \left(1 - \frac{x}{h}\right)\left(1 - \frac{y}{h}\right) f(0, 0) + \left(1 - \frac{y}{h}\right)\frac{x}{h} f(h, 0) + \\ + \left(1 - \frac{x}{h}\right)\frac{y}{h} f(0, h) + \frac{x}{h}\frac{y}{h} f(h, h), \qquad (7.13)$$

точно интерполирует f на всех четырех сторонах квадрата. Более того, Гордон и Холл (1973) показали, что

$$\| D^i (f - \tilde{f}) \|_{\mathscr{L}_\infty (\bar{R})} = O(h^{4-|i|}) \quad (0 \leqslant |i| \leqslant 1),$$

тогда как для простого билинейного интерполянта $\tilde{\tilde{f}}$

$$\| D^i (f - \tilde{\tilde{f}}) \|_{\mathscr{L}_\infty (\bar{R})} = O(h^{2-|i|}) \quad (0 \leqslant |i| \leqslant 1).$$

В качестве численного примера использования смешанных функциональных интерполянтов получим конечноэлементное решение задачи о потенциальном течении в области, представляющей собой единичный квадрат, с источником в точке $x = 0.437$, $y = -k$ $(k > 0)$. Точным решением этой задачи является функция

$$u = \log r,$$

где $r^2 = (x - 0.437)^2 + (y + k)^2$. Поэтому нам нужно найти приближенное решение уравнения

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad ((x, y) \in (0, 1) \times (0, 1)),$$

удовлетворяющее условию $u = \log r$ на границе области.

Разобьем область на N^2 квадратных элементов и применим метод Галеркина для нахождения приближенного решения, являющегося билинейным на не соприкасающихся с границей элементах и включающего билинейные смешанные интерполянты по граничным значениям на элементах, примыкающих к границе. Другими словами, мы ищем приближенное решение вида

$$U(x, y) = W(x, y) + \sum_{j, k=1}^{N-1} U_{jk} \varphi_{jk}(x, y), \qquad (7.14)$$

где φ_{jk} $(j, k = 1, \ldots, N - 1)$ суть кусочные билинейные базисные функции, соответствующие точкам $(j/N, k/N)$, а $W(x, y)$ является кусочной билинейной смешанной функцией, которая отлична от нуля только на примыкающих к границе элементах и на каждом таком элементе представляет собой частный случай общей формы (7.12). Если элемент R со стороной h имеет два внутренних узла, а противоположная им

сторона ($y = 0$) является частью границы, то на R

$$W(x, y) = \left(1 - \frac{y}{h}\right) f(x, 0),$$

так как функция W линейна на трех других сторонах и равна нулю во внутренних узлах. Аналогично этому, если R есть угловой элемент с одним внутренним узлом и две его стороны $x = 0$ и $y = 0$ являются частями границы, то на R

$$W(x, y) = \left(1 - \frac{y}{h}\right) f(x, 0) + \left(1 - \frac{x}{h}\right) f(0, y) - \left(1 - \frac{x}{h}\right)\left(1 - \frac{y}{h}\right) f(0, 0).$$

Для сравнения эта задача решалась также с использованием аппроксимации, которая является билинейной на каждом элементе и в которой граничные условия интерполировались только по значениям в узлах. Численные результаты были получены для 16, 64, 144 и 256 элементов при $k = 0.3$, 0.2 и 0.1. В табл. 4 максимальные по модулю значения приближенных решений, выбранные каждый раз по сетке 16-ти элементов, сравниваются со значениями точного решения в соответствующих точках.

Таблица 4

Число элементов	Граничные условия					
	Дискретные	Точные	Дискретные	Точные	Дискретные	Точные
	$k = 0.3$		$k = 0.2$		$k = 0.1$	
16	2.5747	2.5918	2.7590	2.7852	2.9756	3.0176
64	2.5875	2.5915	2.7822	2.7882	3.0200	3.0310
144	2.5896	2.5914	2.7859	2.7886	3.0281	3.0325
256	2.5904	2.5913	2.7872	2.7888	3.0306	3.0331
Точное решение	2.5913		2.7888		3.0339	

Упражнение 5. Проведите такие же вычисления для задачи с периодическими граничными условиями, когда точным решением является функция

$$u = \sin kxe^{-ky} \quad (k = 2.4),$$

и сравните ваши численные результаты с теми, которые были получены для этой задачи Маршаллом и Митчеллом (1973).

До сих пор мы строили смешанные функциональные интерполянты для прямоугольных элементов. Но они могут быть

построены также и для треугольных элементов, и мы отсылаем интересующегося читателя к работам Барнхилла, Биркгофа и Гордона (1973), Барнхилла и Грегори (1976a) и (1976b) и Маршалла (1975).

7.4. Приложения

(A) Задачи о полях

Теперь мы рассмотрим решение методами конечных элементов некоторых типичных граничных задач о полях.

Задача 1. (Уравнение Пуассона.)

Вернемся к задаче, впервые рассмотренной в гл. 3: найти решение уравнения

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 2$$

в области R, удовлетворяющее граничному условию

$$u = 0,$$

на границе ∂R, где $R = \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. В методе Ритца (м. Р.) минимизируется функционал

$$I(v) = \iint_R \left\{ \frac{1}{2} \left(\left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \right) + 2v \right\} dx\, dy,$$

тогда как в методе наименьших квадратов (м. н. к.) минимизации подлежит функционал

$$J(w) = \iint_R \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} - 2 \right)^2 dx\, dy,$$

где v и w имеют вид (7.1) (или (7.2)) и удовлетворяют граничному условию на ∂R. В модификации Брамбла — Шатца (м. Б. Ш.) метода наименьших квадратов (разд. 5.4(D)) минимизируется функционал

$$J(w) = \iint_R \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} - 2 \right)^2 dx\, dy + h^{-3} \int_{\partial R} w^2\, d\sigma,$$

где весовой множитель h^{-3} введен для того, чтобы оба интеграла имели одинаковую размерность, и уже не требуется, чтобы w удовлетворяла граничному условию на ∂R. Величина h равна значению параметра аппроксимирующего подпространства.

Область R разобьем на квадратные элементы прямыми линиями, параллельными осям x и y, и пусть расстояние между

соседними линиями равно h, где $h = \pi/N$. Треугольные элементы получаются путем проведения в квадратах диагоналей с наклоном —1. Задача 1 была решена численно при $N = 6$ методом Ритца, методом наименьших квадратов и методом Брамбла — Шатца с использованием различных базисных функций. Максимальные ошибки для каждого случая приведены в табл. 5. В общем для метода наименьших квадратов

Таблица 5

Тип базисных функций	м. Р	м. н. к.	м. Б. Ш.
Билинейные на квадрате	0.03269		
Кубические эрмитовы на квадрате	0.00063	0.00063	0.06294
Линейные на треугольнике	0.03116		
Кубические на треугольнике	0.00007	0.00283	0.02935
Пятой степени на треугольнике	0.00041	0.00042	0.00321
Функции Клафа и Точера	0.00036	0.04052	0.03752
Функции Дюпюи и Гёля	0.00032	0.03395	0.03320

и метода Брамбла — Шатца они значительно больше, чем для метода Ритца. По-видимому, это объясняется плохой обусловленностью тех систем линейных уравнений, к которым сводятся методы наименьших квадратов.

Задача 2 (пластина с заделанным краем).

Найти решение уравнения

$$\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = \frac{q}{D}$$

в области R, удовлетворяющее граничным условиям

$$u = \frac{\partial u}{\partial n} = 0$$

на ∂R, где $R = (0, L) \times (0, L)$. Мы рассмотрим тот случай, когда нагрузка q распределена по пластине равномерно. Если $w_{\max}$ есть максимальное смещение пластины, то

$$w_{\max} = \alpha \frac{q}{D} L^4$$

и точное значение α равно 0.00127. Задача решалась только методом Ритца, и вычисленные при $N = 6$ и различных базисных функциях значения α приведены в табл. 6.

Авторы признательны М. Вайну за предоставленные им численные результаты, приведенные здесь в табл. 5 и 6. Дальнейшие подробности и численные результаты для этих и сходных с ними задач можно найти у Вайна (1973).

Таблица 6

Тип базисных функций	
Эрмитовы кубические на квадрате	0.00128
Кубические на треугольнике	0.00136
Пятой степени на треугольнике	0.00127
Функции Клафа и Точера	0.00117
Функции Дюпюи и Гёля	0.00119

В) Задачи о точном управлении для параболических уравнений (Харли и Митчелл, 1976)

Теперь применим метод конечных элементов для решения задачи о точном управлении в случае линейного параболического уравнения. Пусть базисные функции состоят из кусочных бикубических полиномов, а дифференциальное уравнение удовлетворяется на каждом элементе в гауссовых квадратурных точках в смысле метода коллокации.

Рассмотрим уравнение теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \varphi(x, t) \quad ((x, t) \in Q = (0, 1) \times (0, T)), \tag{7.15}$$

дополненное начальным условием

$$u(x, 0) = u_0(x) \quad (x \in [0, 1]) \tag{7.16}$$

и граничными условиями для $t \in [0, T]$, состоящими из условия

$$\frac{\partial u(0, t)}{\partial x} = f(t) \tag{7.17}$$

и либо условия

$$\frac{\partial u(1, t)}{\partial x} = g(t), \tag{7.18a}$$

или же условия

$$\frac{\partial u(1, t)}{\partial x} = \rho [u(1, t) - G(t)], \tag{7.18b}$$

где ρ есть константа. Функции $\varphi(x, t)$ в (7.15) и $f(t)$ в (7.17) заданы, а определить требуется граничную управляющую функцию $g(t)$ (или $G(t)$) так, чтобы решение указанной выше системы в некоторый фиксированный момент времени T *в точности совпало* с $u_d(x)$, т. е. чтобы

$$u(x, T) = u_d(x) \quad (x \in [0, 1]), \tag{7.19}$$

где $u_d(x)$ есть заданная функция.

Область Q нормируем по времени так, чтобы $Q = (0, 1) \times (0, 1)$, и разобьем ее на N^2 квадратов со стороной $h (= 1/N)$, а приближенное решение $U(x, t)$ представим через кусочные бикубические полиномы по x и t (см. разд. 4.2). Полное число коэффициентов в такой аппроксимации равно $4(N + 1)^2$, но, конечно, некоторые из них определяются с помощью функций $u_0(x)$, $u_d(x)$ и $f(t)$, и если число неизвестных коэффициентов есть $M (< 4(N + 1)^2)$, то идеальным для метода коллокации был бы тот случай, когда $U(x, t)$ удовлетворяет дифференциальному уравнению (7.15) в M точках, выбранных в области Q. Это дало бы M линейных уравнений относительно M неизвестных параметров. Будем получать эти уравнения по методу коллокации в квадратурных точках Гаусса для каждого элемента. Если на каждом элементе взять по четыре таких точки, то точность аппроксимации заданными базисными функциями будет наилучшей (ср. с разд. 3.4). При этом получится $4N^2$ уравнений с M неизвестными, т. е. недоопределенная система. Это нежелательно, и поэтому мы возьмем по девять гауссовых узлов на каждом элементе, что даст $9N^2$ уравнений с M неизвестными, т. е. переопределенную систему. Последняя решалась численно по методу наименьших квадратов — с деталями можно познакомиться в работе Харли и Митчелла (1977).

Здесь представлены численные результаты для двух таких задач, точные решения которых известны. Управляющая функция в первой задаче является составной частью граничного условия Неймана (7.18a), а во второй задаче — составной частью смешанного граничного условия (7.18b). В обоих слу-

Таблица 7

	Функция состояния		
N^2	9	16	25
$\% < h^2$	100	100	100
$\% < h^4$	100	89	80

Таблица 8

	Функция состояния			Управляющая функция		
N^2	9	16	25	9	16	25
$\% < h^2$	94	98	92	100	100	100
$\% < h^4$	60	50	40	50	28	20

чаях для функции состояния и функции управления вычислялись модули ошибок в рассматриваемых узлах. В табл. 7 и 8 приводится процент таких узлов от их общего числа, в которых модули ошибок были меньше чем h^2 и h^4.

Задача 1.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad ((x, t) \in Q),$$
$$u_0(x) = \sin x + \cos x,$$
$$u_d(x) = e^{-1}(\sin x + \cos x)$$

и

$$f(t) = e^{-t}.$$

Требуется найти функцию управления $g(t)$ (см. (7.18a)), если точное решение имеет вид

$$u(x, t) = e^{-t}(\sin x + \cos x) \quad ((x, t) \in Q)$$

и

$$g(t) = e^{-t}(\cos(1) - \sin(1)) \quad (t \in [0, 1]).$$

Задача 2.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + e^{-t}\{(4x^2 - 1)\sin(x^2) - 2\cos(x^2)\} \quad ((x, t) \in Q),$$
$$u_0(x) = \sin(x^2),$$
$$u_d(x) = e^{-1}\sin(x^2)$$

и

$$f(t) = 0.$$

Требуется найти управляющую функцию $G(t)$ при $\rho = 2$ (см. (7.18b)), если точное решение имеет вид

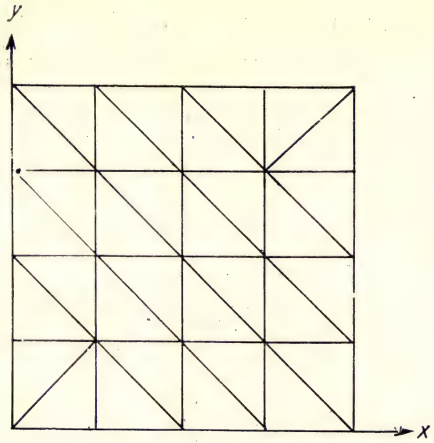
$$u(x, t) = e^{-t}\sin(x^2) \quad ((x, t) \in Q)$$

и

$$G(t) = e^{-t}(\sin(1) - \cos(1)) \quad (t \in [0, 1]).$$

(C) Практическое использование двойственных вариационных принципов

В гл. 2 было приведено несколько примеров таких задач, для которых справедливы двойственные вариационные принципы. Почти во всех случаях численное решение таких задач с помощью метода конечных элементов основывается на принципе минимума, а не на принципе максимума. Вообще говоря, это происходит потому, что принцип минимума легче реализовать практически. Сейчас мы решим с помощью метода конечных элементов одну простую задачу, используя сначала принцип минимума, а затем принцип максимума, и сравним полученные результаты.



$N=25$

Рис. 31.

Выбранная нами задача состоит в нахождении функции u, которая удовлетворяет уравнению

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=u(x,\ y) \quad ((x,\ y)\in R) \tag{7.20}$$

и граничному условию

$$u=g \quad ((x,\ y)\in \partial R), \tag{7.21}$$

где R есть единичный квадрат $0\leqslant x,\ y\leqslant 1$, а функция g задана на границе ∂R этого квадрата. Область разобьем на треугольные элементы, как показано на рис. 31.

Принцип минимума для этой задачи сводится к нахождению

$$\min_u J,$$

где функции u удовлетворяют граничному условию, а

$$J=\frac{1}{2}\iint\limits_R\left\{\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial u}{\partial y}\right)^2+u^2\right\}dx\,dy. \tag{7.22}$$

Если на каждом треугольнике решение аппроксимируется линейным интерполянтом, определяемым значениями u в вершинах треугольника, то полный интерполянт U запишется в виде

$$U(\mathbf{x})=\sum_{i=1}^{N}U_i\varphi_i(\mathbf{x}),$$

где N есть число узлов, а $\varphi_i(\mathbf{x})$ $(i = 1, 2, \ldots, N)$ — кусочно-линейные базисные функции. Неизвестные значения U_i во внутренних узлах получаются путем минимизации (7.22) после замены u на U.

Принцип максимума (Артурс и Ривс, 1974) сводится к нахождению

$$\max_{U_1, U_2} G,$$

где $U_1 = \partial u/\partial x$, $U_2 = \partial u/\partial y$, а

$$G = -\frac{1}{2}\iint_R \left\{\left(\frac{\partial U_1}{\partial x} + \frac{\partial U_2}{\partial y}\right)^2 + U_1^2 + U_2^2\right\} dx\, dy + \\ + \int_{\partial R} \{-U_2 g\, dx + U_1 g\, dy\}. \qquad (7.23)$$

Такая форма принципа максимума с использованием производных первого порядка аналогична принципу минимума дополнительной энергии для задачи о малых упругих деформациях из разд. 2.6. В то же время для аппроксимации U_1 и U_2 на каждом треугольнике используются линейные функции. Неизвестные значения $(U_1)_i$ и $(U_2)_i$ получаются при нахождении максимума функционала (7.23) численным путем.

В качестве примера с помощью принципа минимума и принципа максимума решалась задача, для которой гранич-

Таблица 9

N	J	G
9	4.8872	4.8288
16	4.8563	4.8382
25		4.8416
Точное решение		4.8462

u в середине квадрата

N	Принцип минимума	Принцип максимума
9	2.0268	2.0295
25		2.0279
Точное решение		2.0281

ное условие получается из точного решения

$$u = e^{(x+y)/\sqrt{2}}.$$

Численные результаты приведены в табл. 9. С дальнейшими деталями можно познакомиться в работе Фримана и Гриффитса (1976).

(D) Критическая температура детонирующего стержня

В этом примере полудискретная форма метода конечных элементов (разд. 6.3) применяется для определения критической температуры $\theta_{\text{крит}}$ твердого детонирующего стержня, один конец которого поддерживается холодным ($\theta = \theta_1$), а другой — горячим ($\theta = \theta_0$). Критическая температура определяется так, что при $\theta_0 < \theta_{\text{крит}}$ мы со временем придем к стационарному решению $\theta \leqslant \theta_0$ на всем стержне, тогда как при $\theta_0 > \theta_{\text{крит}}$ в некоторой внутренней точке стержня со временем возникнет $\theta > \theta_0$, что и является признаком зажигания стержня.

Уравнение, описывающее предшествующий зажиганию процесс, может быть записано в безразмерных единицах (Кук, 1958) как

$$\frac{\partial\theta}{\partial t} = \frac{\partial^2\theta}{\partial x^2} + C\exp\left(r - \frac{1}{\theta}\right) \quad (0 < x < 1,\ t > 0),$$

где C и r являются константами; оно дополняется начальным условием

$$\theta(x,\ 0) = \theta_1 \quad (0 < x < 1)$$

и граничными условиями

$$\theta(0,\ t) = \theta_0$$

и

$$\theta(1,\ t) = \theta_1,$$

где $\theta_0 > \theta_1$.

Задача сведется к системе нелинейных обыкновенных дифференциальных уравнений, если приближенное решение ищется в виде

$$\theta(x,\ t) = \sum_{i=0}^{N+1} \alpha_i(t)\,\varphi_i(x),$$

где базисные функции φ_i ($i = 0, 1, \ldots, N+1$) определены на интервале $[0, 1]$ с помощью равномерного разбиения. Функции α_0 и α_{N+1} определяются граничными условиями, а $\alpha_i(t)$ ($i = 1, 2, \ldots, N$) удовлетворяют системе уравнений

$$A_1\dot{\alpha} = -A_2\alpha + \mathbf{f}(\alpha), \tag{7.24}$$

где A_1 и A_2 есть положительно определенные матрицы, точка означает дифференцирование по времени, а $\mathbf{f}(\alpha)$ есть нелинейная векторная функция, компоненты которой имеют вид

$$f_j(\alpha) = \int_0^1 \varphi_j(x)\, C \exp\left(r - \left(\sum_{i=0}^{N+1} \alpha_i(t)\, \varphi_i(x)\right)^{-1}\right) dx \quad (j = 1, \ldots, N).$$

Система (7.24) является *жесткой*, и при ее численном решении необходимо проявлять осмотрительность. Она решалась с переменным по времени шагом методом, предложенным Т. Р. Гопкинсом, которому авторы признательны за предоставленные им численные результаты.

В численном примере полагалось, что решение станет стационарным, если $\theta(x, 10) < \theta_0$ для $x \in (0, 1)$; напротив, если $\theta(x, t) \geqslant \theta_0$ для некоторого $x \in (0, 1)$ и некоторого $t \leqslant 10$, то считалось, что произошло зажигание. Критическая температура находилась по методу деления отрезка пополам: если $\theta_0 = \theta_L$ приводит к стационарному решению, а $\theta_0 = \theta_H$ $(> \theta_L)$ приводит к зажиганию, то вычисляется решение для $\theta_0 =$

Таблица 10

Линейные базисные функции

θ_0	0.02440	0.03050	0.02745	0.02898	0.02821	0.02859
Время зажигания	∞	0.90	∞	4.74	∞	6.04
θ_0	0.02840	0.02850	0.02855	0.02857	0.02858	
Время зажигания	∞	∞	∞	∞	6.5	

Таблица 11

h	1/4	1/8	1/16	1/32
Линейные: $h = \frac{1}{N+1}$		0.02857	0.02843	0.02839
Квадратичные: $h = \frac{2}{N+1}$	0.02837	0.02841	0.02838	
Эрмитовы кубические: $h = \frac{2}{N}$	0.02822	0.02837	0.02837	

$= \frac{1}{2}(\theta_L + \theta_H)$ и θ_H или θ_L заменяется этим новым значением θ_0 в зависимости от того, приводит такое θ_0 к зажиганию или нет.

В табл. 10 и 11 приведены численные результаты, полученные при решении этой задачи с различными значениями N для кусочно-линейных, кусочно-квадратичных и эрмитовых кусочно-кубических базисных функций. Во всех случаях $\theta_1 = 0.0122$, что примерно соответствует температуре 12 °C, а бесконечное время зажигания означает то, что решение становится стационарным. Кусочно-квадратичные базисные функции изображены на рис. 33.

(E) Задачи о конвекционной проводимости

На неадекватность многих стандартных конечноэлементных методов в применении к задачам, содержащим как первые, так и вторые производные искомой функции, впервые обратил внимание авторов О. Зенкевич. Это в особенности так, если коэффициенты при первых производных оказываются сравнительно большими. Типичным примером такого рода служит задача о стационарном течении несжимаемой вязкой жидкости; здесь уравнение переноса завихрения в двумерном случае имеет вид

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} - \frac{1}{\nu}\left(u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y}\right) = 0, \tag{7.25}$$

где w обозначает завихрение, u и v являются компонентами скорости, а ν есть коэффициент кинематической вязкости. Коэффициенты при первых производных в (7.25) по порядку величины эквивалентны числу Рейнольдса и поэтому принимают большие значения во многих практических задачах.

Чтобы показать те трудности, с которыми приходится сталкиваться при численном решении уравнения (7.25), рассмотрим одномерное модельное уравнение

$$\frac{d^2 w}{dx^2} - k\frac{dw}{dx} = 0 \quad (x \in [0, 1]), \tag{7.26}$$

где $k = \mu/\nu$ положительно и предполагается постоянным. Разобьем интервал $[0, 1]$ на N равных частей длины $h = 1/N$ точками $x = ih$ $(i = 0, 1, \ldots, N)$. В численных примерах использовались граничные условия

$$w = \begin{cases} 1 & (x = 0), \\ 0 & (x = 1), \end{cases} \tag{7.27}$$

получающиеся из точного решения

$$w(x) = \frac{e^k - e^{kx}}{e^k - 1}. \tag{7.28}$$

Решение Галеркина W удовлетворяет системе уравнений

$$(W', \varphi_i') + k(W', \varphi_i) = 0 \quad (i = 1, 2, \ldots, N-1), \tag{7.29}$$

где

$$W(x) = \sum_{i=0}^{N} W_i \varphi_i(x) \tag{7.30}$$

и штрих означает дифференцирование по x. Если предположить базисные функции φ_i кусочно-линейными, то (7.29) сведется к системе разностных уравнений

$$\left(1 - \tfrac{1}{2}kh\right) W_{i+1} - 2W_i + \left(1 + \tfrac{1}{2}kh\right) W_{i-1} = 0 \tag{7.31}$$

$$(i = 1, 2, \ldots, N-1).$$

Эта система имеет точное решение

$$W_i = A_1 + B_1 \left(\frac{1 + \frac{1}{2}kh}{1 - \frac{1}{2}kh}\right)^i \quad (i = 0, 1, \ldots, N),$$

и поэтому при

$$h > \frac{2}{k}$$

будут иметь место осцилляции.
Теперь заменим (7.29) системой

$$(W', \psi_i') + k(W', \psi_i) = 0 \quad (i = 1, 2, \ldots, N-1),$$

где W снова имеет вид (7.30) с теми же самыми кусочно-линейными базисными функциями φ_i, а в качестве пробных функций ψ_i $(i = 1, 2, \ldots, N-1)$ взяты асимметричные кусочно-линейные базисные функции, показанные на рис. 32. Такие функции были предложены Д. Ф. Гриффитсом и имеют вид

$$\psi_i(x) = \begin{cases} \frac{1+\alpha\eta}{1-\eta}\left(1 + \frac{x}{h} - i\right) & (x \in [(i-1)h, (i-\eta)h]), \\ 1 + \alpha\left(i - \frac{x}{h}\right) & (x \in [(i-\eta)h, (i+\xi)h]), \\ \frac{1-\alpha\xi}{\xi-1}\left(\frac{x}{h} - 1 - i\right) & (x \in [(i+\xi)h, (i+1)h]), \end{cases} \tag{7.32}$$

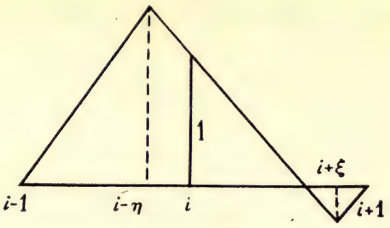


Рис. 32.

где $0 \leqslant \xi, \eta < 1$ и α есть тангенс угла наклона среднего звена со знаком минус. Асимметричные функции ψ_i совпадают со стандартными кусочно-линейными базисными функциями φ_i при $\alpha = \xi = \eta = 0$. Теперь коэффициенты аппроксимации (7.30) будут удовлетворять системе разностных уравнений

$$\left[1 - \tfrac{1}{2} kh(1 + \xi(1-\alpha))\right] W_{i+1} - \\ - \left[2 + \tfrac{1}{2} kh(\eta(\alpha+1) + \xi(\alpha-1))\right] W_i + \\ + \left[1 + \tfrac{1}{2} kh(1 + \eta(1+\alpha))\right] W_{i-1} = 0 \quad (i = 1, 2, \ldots, N-1), \tag{7.33}$$

имеющей точное решение

$$W_i = A_2 + B_2 \left[\frac{1 + \frac{1}{2} kh + \frac{1}{2} kh\eta(1+\alpha)}{1 - \frac{1}{2} kh - \frac{1}{2} kh\xi(1-\alpha)}\right]^i \quad (i = 0, 1, 2, \ldots, N). \tag{7.34}$$

Разностные уравнения (7.33) имеют первый порядок точности, если $\xi \neq \eta$ и

$$\alpha = \frac{\xi + \eta}{\xi - \eta}.$$

Если ввести еще обозначение

$$A = \frac{2\xi\eta}{\xi - \eta},$$

то точное решение (7.34) перепишется в виде

$$W_i = A_2 + B_2 \left[\frac{1 + \frac{1}{2} kh + \frac{1}{2} Akh}{1 - \frac{1}{2} kh + \frac{1}{2} Akh}\right]^i.$$

Следовательно, в конечноэлементном решении не будет осцилляций, если

(I) $A \geqslant 1$

или

(II) $-\infty < A < 1$ и $h < \frac{2}{(1-A)k}$.

Отметим, что три стандартные конечноразностные аппроксимации уравнения (7.26) получаются в следующих случаях:

(I) $A = 1$ (разность назад),
(II) $A = 0$ (центральная разность),
(III) $A = -1$ (разность вперед).

При $A = 0$ получаются уравнения Галеркина (7.31); только они имеют второй порядок точности.

Наконец, мы получим решение Галеркина, используя кусочные квадратичные базисные функции, показанные на рис. 33. После проведения соответствующих выкладок получаются разностные уравнения

$$\left(1 - \tfrac{1}{2}kh\right) W_{i+1} - 4\left(2 - \tfrac{1}{2}kh\right) W_{i+1/2} + 14W_i - $$
$$- 4\left(2 + \tfrac{1}{2}kh\right) W_{i-1/2} + \left(1 + \tfrac{1}{2}kh\right) W_{i-1} = 0$$
$$(i = 1, 2, \ldots, N-1)$$

в целых узлах и

$$(4 - kh) W_i - 8W_{i-1/2} + (4 + kh) W_{i-1} = 0 \quad (i = 1, 2, \ldots, N)$$

в полуцелых узлах. После исключения неизвестных в полуцелых узлах получим систему разностных уравнений

$$\left(1 - \tfrac{1}{2}kh + \tfrac{1}{12}k^2h^2\right) W_{i+1} - \left(2 + \tfrac{1}{6}k^2h^2\right) W_i +$$
$$+ \left(1 + \tfrac{1}{2}kh + \tfrac{1}{12}k^2h^2\right) W_{i-1} = 0 \quad (i = 1, 2, \ldots, N-1),$$

имеющую точное решение

$$W_i = A_3 + B_3 \left[\frac{1 + \frac{1}{2}kh + \frac{1}{12}k^2h^2}{1 - \frac{1}{2}kh + \frac{1}{12}k^2h^2}\right]^i.$$

Это решение свободно от осцилляций при любых значениях h и k.

Численные результаты приведены в табл. 12, где сравниваются конечноэлементные решения, полученные для линейных и квадратичных базисных функций. Для линейных функций были выбраны два случая, а именно $A = 0$ (центральные

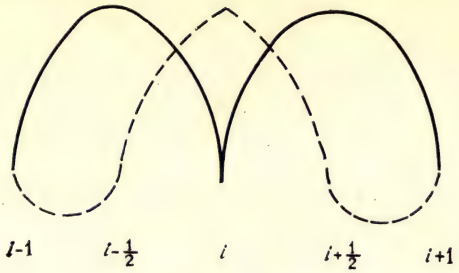


Рис. 33.

разности) и $A = 1$ (разности назад). Осцилляции видны в случае центральных разностей при $h = 1/20$. Не вызывает сомнения то, что для этой частной задачи можно получить лучшую точность при том же объеме вычислений, уменьшая размеры элементов с приближением к правой границе. Однако в тех случаях, когда форма ответа не известна заранее, неравномерные сетки редко используются на первой стадии работы над задачей. Дальнейшие подробности по этому вопросу можно найти у Кристи (1975).

(F) Сингулярные изопараметрические элементы

В разд. 4.6 было показано, что необходимо соблюдать осторожность при выборе узлов на изопараметрических элементах, чтобы нигде на элементе якобиан не обращался в нуль. Однако бывают такие ситуации, при которых обращение якобиана в нуль в отдельной точке может быть полезным. В этом разделе мы рассмотрим вкратце один случай применения таких изопараметрических элементов.

Если функция $u(x, y)$ удовлетворяет уравнению

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

в области R, то в *окрестности угловой точки* границы ∂R она может быть представлена в виде

$$u = \sum_{j=1}^{\infty} \gamma_j r^{j/\alpha} \sin\left(\frac{j\theta}{\alpha}\right) \tag{7.35}$$

при некоторых постоянных γ_j $(j = 1, \ldots)$, где $\alpha\pi$ есть величина угла, внутри которого лежит R, а (r, θ) есть полярные координаты с началом в вершине угла. Отсюда следует, что вблизи вершины врезающегося в область угла $(\alpha > 1)$ производные главного члена в (7.35) неограниченно возрастают

Таблица 12

($k = 60$)

$h = 1/20$				
x	Точное решение	Решение для центральной разности	Решение для разности назад	Решение для квадратичных функций
0.90	0.9975	0.9600	0.9375	0.9941
0.95	0.9502	1.2000	0.7500	0.9231
1.00	0	0	0	0

$h = 1/40$				
x	Точное решение	Решение для центральной разности	Решение для разности назад	Решение для квадратичных функций
0.90	0.9975	0.9996	0.9744	0.9974
0.925	0.9889	0.9971	0.9360	0.9885
0.95	0.9502	0.9796	0.8400	0.9490
0.975	0.7769	0.8571	0.6000	0.7742
1.00	0	0	0	0

при стремлении r к нулю. Следующие две ситуации имеют очень много общего: когда $\alpha = 2$, т. е. область имеет разрез или трещину, и когда $\alpha = {}^3/_2$, т. е. область имеет прямоугольную «вмятину». Тогда главные члены разложения становятся пропорциональными $r^{1/2}$ и $r^{2/3}$ соответственно. Одна из основных причин непригодности многих стандартных численных методов для решения таких задач состоит в том, что эти функции нельзя с достаточной точностью приблизить полиномами (по r).

Теперь мы приведем два примера таких изопараметрических элементов, которые позволяют обойти эту трудность при условии, что угловая точка области является вершиной элемента, в котором другие узлы выбраны специальным образом.

(1) Для приближения $r^{1/2}$ могут быть использованы *квадратичные элементы.* Если $t_4 = \frac{1}{2}(t_1 + t_2)$, $t_5 = \frac{1}{4}(3t_3 + t_2)$ и $t_6 = \frac{1}{4}(3t_3 + t_1)$ (см. рис. 17), то изопараметрическое преобразование примет вид

$$t - t_3 = ((t_1 - t_3)\,p + (t_2 - t_3)\,q)\,(p + q) \quad (t = x,\ y),$$

и линейные по p и q функции будут иметь нужное поведение вида $r^{1/2}$, где r есть расстояние до вершины P_3. Например,

вдоль P_1P_3 $(q = 0)$

$$r^2 = (x - x_3)^2 + (y - y_3)^2 = ((x_1 - x_3)^2 + (y_1 - y_3)^2)\, p^4,$$

и поэтому $p \approx r^{1/2}$. Якобиан этого преобразования можно представить в виде

$$2C_{123}(p + q)^2,$$

и поэтому он обращается в нуль только в точке P_3 $(p = q = 0)$.

(2) Для приближения $r^{2/3}$ в окрестности узла P_3 могут быть использованы *кубические элементы*. Если узлы расположены правильно, то изопараметрическое преобразование примет вид

$$t - t_3 = ((t_1 - t_3)\, p + (t_2 - t_3)\, q)\,(p + q)^2 \quad (t = x,\ y).$$

По аналогии с предыдущим случаем можно убедиться, что линейные по p и q функции ведут себя как $r^{1/3}$, так что квадратичные функции будут иметь нужное поведение вида $r^{2/3}$. Теперь якобиан преобразования запишется в виде

$$3C_{123}(p + q)^4,$$

и обратится в нуль только при $p = q = 0$.

Дополнительные подробности о сингулярных изопараметрических элементах, иллюстрацию их эффективности при проведении практических вычислений и обобщения такого подхода на различные особенности и на случаи более высоких размерностей можно найти в работе Уэйта (1976).

СПИСОК ЛИТЕРАТУРЫ [1])

Агмон (Agmon S.)
(1965) Lectures on Elliptic Boundary Value Problems, Van Nostrand, Princeton.
Адини, Клаф (Adini A., Clough R. W.)
(1961) Analysis of Plate Bending by the Finite Element Method, Nat. Sci. Found. Rept. G7337. Univ. of California, Berkeley.
Азиз (Aziz A. K. Ed.)
(1972) The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York.
Айронс (Irons B. M.)
(1966) Conf. on Use of Digital Computers in Structural Engineering, Newcastle.
(1969) Int. J. Num. Meth. Eng. 1, 29.
Айронс, Раззак (Irons B. M., Razzaque A.)
(1972) см. Азиз (1972), 557—587.
Артурс (Arthurs A. M.)
(1970) Complementary Variational Principles, Clarendon Press Oxford.
Артурс, Ривс (Arthurs A. M., Reeves R. I.)
(1974) J. Inst. Math. Applics. 14, 1.
Бабушка (Babuska I.)
(1969) Tech. Note BN-624, University of Maryland.
(1971) SIAM J. Numer. Anal., 8, 304.
(1973) Numer. Math., 20, 179.
Бабушка, Азиз (Babuska I., Aziz A. K.)
(1976) SIAM J. Numer. Anal., 13, 214.
Бабушка, Зламал (Babuska I., Zlamal M.)
(1973) SIAM J. Numer. Anal., 10, 863.
Барнхилл, Биркгоф, Гордон (Barnhill R. E., Birkhoff G., Gordon W.)
(1973) J. Approx. Theory, 8, 114.
Барнхилл, Грегори, Уайтман (Barnhill R. E., Gregory A., Whiteman J. R.)
(1972) см. Азиз (1972), 749—755.
Барнхилл, Грегори (Barnhill R. E., Gregory A.)
(1976a) Math. Comp. (to appear).
(1976b) J. Approx. Theory (to appear).
Бейкер (Baker G. A.)
(1973) Math. Comp., 27, 229.
Березин И. С., Жидков Н. П.
(1962) Методы вычислений, т. 2. — М.: Наука, 1962.
Бергер (Berger A. E.)
(1972) см. Азиз (1972), 757—796.
(1973) Numer. Math., 21, 345.

---

[1]) При цитировании литературы номера страниц указываются по русскому переводу (если такой имеется), но указывается год выхода в свет оригинала.

Бергер, Скотт, Стренг (Berger A. E., Scott R., Strang G.)
(1972) Symposia Mathematica X. Academic Press, London.
Берс, Джон, Шехтер (Bers L., John F., Schechter M.)
(1964) Partial Differential Equations, Interscience New York. (Русский перевод: Берс Л., Джон Ф., Шехтер М. Уравнения с частными производными. — М.: Мир, 1966.)
Биркгоф, Шульц, Варга (Birkhoff G., Schultz M. H., Varga R. S.)
(1968) Numer. Math., 11, 232.
Биркгоф (Birkhoff G.)
(1971) Proc. Nat. Acad. Sci., 68, 1162.
Биркгоф, Менсфилд (Birkhoff G., Mansfield L.)
(1974) J. Math. Anal. Applics, 47, 531.
Бонд, Свенелл, Геншелл, Уабартон (Bond T. J., Swanell R. D., Henshell R. D., Warburton G. B.)
(1973) J. Strain Anal., 8, 182.
де Бур (de Boor C.)
(1974) Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York.
де Бур, Шварц (de Boor C., Swartz B.)
(1973) SIAM J. Numer. Anal., 10, 582.
Брамбл, Дюпон, Томе (Bramble J. H., Dupont I., Thomee V.)
(1972) Math. Comp., 26, 869.
Брамбл, Гильберт (Bramble J. H., Hilbert S. R.)
(1970) SIAM J. Numer. Anal., 7, 112.
Брамбл, Зламал (Bramble J. H., Zlamal M.)
(1970) Math. Comp., 24, 809.
Брамбл, Нитше (Bramble J. H., Nitsche J. A.)
(1973) SIAM J. Numer. Anal., 10, 81.
Брамбл, Шатц (Bramble J. H., Schatz A. H.)
(1970) Comm. P. Appld. Math., 23, 635.
(1971) Math. Comp., 25, 1.
Брамбл, Томе (Bramble J. H., Thomee V.)
(1974) R. A. I. R. O., 8 (R-2), 5.
Браун (Brown J. H.)
(1975) Non Comforming Finite Elements and their Applications, M. Sc. Thesis, Univ. of Dundee.
Вайн (Vine M.)
(1973) Applications of the Finite Element Method to Partial Differential Equations, Ph. D. Thesis, Univ. of Dundee.
Вайнберг М. М.
(1956) Вариационные методы исследования нелинейных операторов. М., ГИТТЛ.
Варга (Varga R. S.)
(1971) Functional Analysis and Approximation Theory in Numerical Analysis. SIAM Publication Philadelphia.
(Русский перевод: Варга Р. Функциональный анализ и теория аппроксимации в численном анализе. — М.: Мир, 1974.)
Васидзу (Washizu K.) [1])
(1968) Variationl Methods in Elasticity and Plasticity, Pergamon Press, London.
Ватсон (Watson G. A., Ed.)
(1974) Conf. Num. Soln. Diff. Equns, Dundee, Springer-Verlag Lecture Notes in Maths., 363.

[1]) Транскрибируется также как Вашицу.

Ватсон (Watson G. A. Ed.)
(1976) Conf. Num. Anal., Dundee, Springer-Verlag Lecture Notes in Maths.
Вильсон, Тейлор, Доерти, Габусси (Wilson E. L., Taylor R. L., Doherty W. P., Ghaboussi J.)
(1971) Univ. of Illinois Symposium.
Вулих Б. З.
(1967) Введение в функциональный анализ. — М.: Наука.
Гопкинс, Уэйт (Hopkins T. R., Wait R.)
(1976) Comp. Meth. App. Mech. Eng., 9, 181.
Гордон (Gordon W. J.)
(1971) SIAM J. Numer. Anal., 8, 158.
Гордон, Уиксом (Gordon W. J., Wixom J. A.)
(1974) SIAM J. Numer. Anal., 11, 909.
Гордон, Холл (Gordon W. J., Hall C. A.)
(1973) Numer. Math., 21, 109.
Грам (Gram J. G.)
(1973) Numerical Solution of Partial Differential Equations, Reidel Publishing Co., Boston, U. S. A.
Демьянович Ю. К.
(1964) Мат. сб., 5, 1452.
Денди (Dendy J. E.)
(1975) SIAM J. Numer. Anal., 12, 541.
Денди, Файервезер (Dendy J. E., Fairweather G.)
(1975) SIAM, J. Numer. Anal., 12, 144.
Джордан (Jordan W. B.)
(1970) A. E. C. Research and Development Report KAPL-M-7112.
Дуглас, Дюпон (Douglas J., Dupont T.)
(1970) SIAM J. Numer. Anal., 7, 575.
(1971) см. Хаббард (1971), 133—244.
(1973) Math. Comp., 27, 17.
(1975) Math. Comp., 29, 360.
Дуглас, Дюпон, Уилер (Douglas J., Dupont T., Wheeler M. F.)
(1974) R. A. I. R. O., 8 (R-2), 61.
Дюпон, Файервезер, Джонсон (Dupont T., Fairweather G., Johnson J. P.)
(1974) SIAM J. Numer. Anal., 11, 392.
Дюпюи, Гёль (Dupuis G., Göel J. J.)
(1970) Int. J. Num. Meth. Eng., 2, 563.
Девис, Рабинович (Davis P. J., Rabinowitz R.)
(1967) Numerical Integration, Blaisdell, Waltham, Mass.
Зенкевич (Zienkiewicz O. C.)
(1967 The Finite Element Method in Structural and Continuum Mechanics, Mc Graw-Hill, New York.
(Русский перевод: Зенкевич О., Чанг И. Метод конечных элементов в теории сооружений и в механике сплошных сред. — М.: Недра, 1974.)
(1971) The Finite Element Method in Engeneering Science, McGraw-Hill,
Зенкевич (Zienkiewicz O. C.)
New York.
(Русский перевод: Зенкевич О. Метод конечных элементов в технике. — М.: Мир, 1975.)
Зламал (Zlamal M.)
(1973) SIAM J. Numer. Anal., 10, 227.
(1974) SIAM J. Numer. Anal., 11, 347.
(1975) Math. Comp., 29, 350.
Иосида (Iosida K.)
(1965) Functional Analysis, Springer-Verlag, Berlin.
(Русский перевод: Иосида К. Функциональный анализ. — М.: Мир, 1967.)

Канторович Л. В.
(1933) Изв. АН СССР, 5, 647.
Клаф, Точер (Clough R. W., Tocher J. L.)
(1965) Proc. Ist. Conf. Matrix. Methods in Structural Mechanics, Wright-Patterson A. F. B., Ohio.
Клегг (Clegg J. C.)
(1967) Calculus of Variations, Oliver and Boyd, Edinburgh.
Комини, дель Гвидичи, Левис, Зенкевич (Comini G., del Guidici S., Lewis R. W., Zienkiewicz O. C.)
(1974) Int. J. Num. Meth. Eng., 8, 613.
Кристи (Christie I.)
(1975) Conduction — Convection Problems. M. Sc. Thesis. Univ. of Dundee.
Крузей, Равьяр (Crouzeix M., Raviart P. A.)
(1973) R. A. I. R. O., 7, (R-3), 33.
Кук (Cook G. B.)
(1958) Proc. Roy. Soc. London. A. 264, 154.
Курант, Гильберт (Courant R., Hilbert D.)
(1953) Methods of Mathematical Physics, Vol. 1, Interscience, New York. (Русский перевод издания 1930 года: Курант Р., Гильберт Д. Методы математической физики, т. 1. — М. — Л.: ГИТТЛ, 1951.)
Ламберт (Lambert J. D.)
(1973) Computational Methods in Ordinary Differential Equations, Wiley, London.
Ланкастер (Lancaster P., Ed.)
(1973) Proc. Conf. on Theory and Applications of Finite Element Methods, Univ. of Calgary.
Ласко, Лесен (Lascaux P., Lesaint P.)
(1975) R. A. I. R. O., 9 (R-1), 9.
Лаури (Laurie D. P.)
(1977) Inst. Math. Applics, 19, 119.
Лионс, Мадженес (Lions J. L., Magenes E.)
(1972) Non-Homogeneous Bondary Value Problems and Applications I, Springer-Verlag, Berlin.
(Русский перевод с французского издания 1968 года: Лионс Ж.-Л., Мадженес Е. Неоднородные граничные задачи и их приложения. — М.: Мир, 1971.)
Лукас, Редин (Lucas T. R., Reddien G. W.)
(1972) SIAM, J. Numer. Anal., 9, 341.
Маклеод (McLeod R. J.)
(1977) J. Approx, Th., 19, 25.
Маклеод, Митчелл (McLeod R. J., Mitchell A. R.)
(1972) J. Inst. Math. Applics, 10, 382.
Маклеод, Митчелл (McLeod R. J., Mitchell A. R.)
(1975) J. Inst. Math. Applics., 16, 239.
Маршалл (Marshall J. A.)
(1975) Some Applications of Blending Function Techniques to Finite Element Methods, Ph. D. Thesis. Univ. of Dundee.
Маршалл, Митчелл (Marshall J. A., Mitchell A. R.)
(1973) J. Inst. Math. Applics, 12, 355.
Миллер (Miller J. J. H., Ed.)
(1973) Topics in Numerical Analysis I., Academic Press, New York.
(1975) Topics in Numerical Analysis II, Academic Press, New York.
Митчелл (Mitchell A. R.)
(1969) Computational Methods in Partial Differential Equations, Wiley, London.
Михлин С. Г.
(1976) Вариационные методы в математической физике. — М.: Наука.

Михлин С. Г., Смолицкий Х. Л.
(1965) Приближенные методы решения дифференциальных и интегральных уравнений. — М.: Наука.
Морс, Фешбах (Morse P. M., Feshbach H.)
(1953) Methods of Theorethical Physics, McGraw-Hill, New York.
(Русский перевод: Морс Р. М., Фешбах Г., Методы теоретической физики. — М.: ИЛ, т. 1, 1958, т. 2, 1960.)
Нечас (Necas J.)
(1967) Les Methodes Directes en Theorie des Equations Elliptiques, Academia, Prague.
Нитше (Nitsche J. A.)
(1971) Abhandt. d. Hamb. Math. Sem. 36, 9.
(1972) см. Азиз (1972, 603—627).
Нитше, Шатц (Nitsche J. A., Schatz A. H.)
(1974) Math. Comp., 28, 937.
Нобл (Noble B.)
(1973) см. Уайтман (1973), 143—152.
Нобл, Севелл (Noble B., Sewell M. J.)
(1972) J. Inst. Math. Applics, 9, 123.
Обэн (Aubin J. P.)
(1972) Approximation of Elliptic Boundary Value Problems, Wiley, New York.
(Русский перевод: Обэн Ж.-П. Приближенное решение эллиптических краевых задач. — М.: Мир, 1977.)
Оганесян Л. А.
(1966) ЖВМ и МФ, 6, 116.
Оганесян Л. А., Руховец Л. А.
(1969) ЖВМ и МФ, 9, 153.
Оден (Oden J. T.)
(1972) Finite Elements of Nonlinear Continua, McGraw-Hill, New York.
(Русский перевод: Оден Дж. Конечные элементы в нелинейной механике сплошных сред. — М.: Мир, 1976.)
Оден, Зенкевич, Галлагер, Тейлор (Oden J. T., Zienkiewicz O. C., Gallagher R. H., Taylor C.) (Eds.)
(1974) Finite Element Methods in Flow Problems, Wiley, New York.
Пиан (Pian T. H. H.)
(1970) Numerical Solution of Field Problems in Continuum Physics, SIAM — AMS Proceedings Volume 2.
Пауэлл (Powell M. J. D.)
(1973) Conference on Numerical Software, Loughborough.
Прентер (Prenter P. M.)
(1975) Splines and Variational Methods, Wiley, New York
Розен (Rosen P.)
(1953) J. Chem. Phys., 21, 1220.
Розен (Rosen P.)
(1954) J. App. Phys., 25, 336.
Севелл (Sewell M. J.)
(1969) Phill. Roy. Soc. (London), A 265, 319.
Семенич, Глэдвелл (Siemenuich J. L., Gladweell I.)
(1974) Numer. Anal. Report 5, Manchester University.
Сербин (Serbin S. M.)
(1975) Math. Comp., 29, 777.
Симмонс (Simmons G. F.)
(1963) Introduction to Topology and Modern Analysis, McGraw-Hill, New York.
Синж (Synge J. L.)
(1957) The Hypercircle in Mathematical Physics, C. U. P., London.

Скотт (Scott R.)
(1975) SIAM J. Numer. Anal., 12, 404.
Стренг (Strang G.)
(1972) см. Азиз (1972), 489—710.
Стренг, Бергер (Strang G., Berger A. R.)
(1971) см. Proc, American Math. Soc. Summer Inst. in Partial Diff. Equns, 199—205.
Стренг, Фикс (Strang G., Fix G.)
(1973) An Analysis of the Finite Element Method, Prentece Hall, New Jersey. (Русский перевод: Стренг Г., Фикс Дж. Теория метода конечных элементов. — М.: Мир, 1977.)
Сьярле (Ciarlet P. G.)
(1973a) Springer-Verlag Lecture Notes, 363, 21, Berlin.
(1973b) см. Уайтман (1973), 113—129.
Сьярле, Равьяр (Ciarlet P. G., Raviart P. A.)
(1972a) Arch. Rat. Mech. Anal., 46, 177.
(1972b) Comp. Meth. Appl. Mech. Eng., 1, 217.
(1972c) см. Азиз (1972), 409—474.
Табаррок (Tabarrok B.)
(1973) Proc. Conf. on Theory and Applications of Finite Element Methods, Univ. of Calgary (Ed. P. Lancaster (1973))
Томе (Thomee U.)
(1973) J. Inst. Math. Applics., 11, 33.
Томе, Уолбин (Thomee V., Wahlbin L.)
(1975) SIAM J. Numer. Anal., 12, 378.
Уайтман (Whiteman J. R., Ed.)
(1973) The Mathematics of Finite Elements and Applications, Academic Press, New York.
Уайтман (Whiteman J. R.)
(1975) A Bibliography for Finite Elements, Academic Press, London.
(1976) The Mathematics of Finite Elements and Applications, Academic Press, New York.
Уачспресс (Wachspress E. L.)
(1971) Conf. on Appl. Num. Anal., Dundee, Springer-Verlag Lecture Notes in Math., 228, 223.
(1973) J. Inst. Math. Applics, 11, 83.
(1974) Conf. Num. Soln. Diff. Equns., Dundee, Springer-Verlag Lecture Notes in Math., 363, 177.
(1975) A Rational Finite Element Basis, Academic Press, New York.
Уилер (Wheeler M. F.)
(1973) SIAM J. Numer. Anal., 10, 723.
Уилкинсон (Wilkinson J. H.)
(1965) The Algebraic Eigenvalue Problem. O. U. P., Oxford. (Русский перевод: Уилкинсон Дж. Х., Алгебраическая проблема собственных значений. — М.: Наука, 1970.)
Уэйт (Wait R.)
(1976) Singular Isoparametric Finite Elements, J. Inst Math. Applics, 20, 133.
Уэйт, Митчелл (Wait R., Mitchell A. R.)
(1971) J. Inst. Math. Applics, 4, 241.
Файервезер (Fairweather G.)
(1972) Galerkin Methods for Differential Equations, CSIR Special Report WISK96, Pretoria, South Africa.
Файервезер, Джонсон (Fairweather G., Johnson J. P.)
(1975) Numer. Math., 23, 269.

Фикс (Fix G. L.)
(1972) см. Азиз (1972), 525—556.
Финлейсон, Скривен (Finlayson B. A., Scriven L. E.)
(1967) Int. J. Heat. Mass. Trans., 10, 799.
Фриман, Гриффитс (Freeman L., Griffiths D. F.)
(1976) Complementary Variational Principles and the Finite Element Method (to appear).
Хаббард (Hubbard B., Ed.)
(1971) Numerical Solution of Partial Differential Equations. II, SYNSPADE, 1970, Academic Press, New York,
Халм (Hulme B. L.)
(1972) Math. Comp., 26, 415.
Харли, Митчелл (Harley P. J., Mitchell A. R.)
(1976) J. Inst. Math. Applics, 18, 9.
(1977) Int. J. Num. Meth.Eng., 11, 345.
Хеболд, Варга (Herbold R. J., Varga R. S.)
(1972) Aeq. Math., 7, 36.
Хильдебранд (Hildebrand F. B.)
(1965) Methods of Applied Mathematics, Prentice Hall, New York.
Чекки, Челла (Cecchi M. M., Cella A.)
(1973) Proc. 4th Canadian Congress on Appld. Mech. 767—768.
Чернука, Купер, Линдберг, Олсон (Chernuka M. W., Cowper G. R., Lindberg G. M., Olson M. D.)
(1972) Int. J. Num. Meth. Eng., 4, 49.
Шехтер (Schechter R. S.)
(1967) The Variational Method in Engeneering, McGraw-Hill, New York. (Русский перевод: Шехтер Р. С. Вариационный метод в инженерных расчетах. — М.: Мир, 1971.)
Шёнберг (Schoenberg I. J.)
(1969) Approximations With Special Emphasis on Spline Functions, Academics Press, New York.
Элберг, Ито (Ahlberg J. H., Ito T.)
(1975) Math. Comp., 29, 761.
Эльсгольц Л. Э.
(1958) Вариационное исчисление. М., ГИТТЛ.
Эргатодис, Айронс, Зенкевич (Ergatoudis I., Irons B. M., Zienkiewicz O. C.)
(1968) Int. J. Solids Structures, 4, 31.

Список дополнительной литературы

1. С математической постановкой и теоретическими аспектами большей части рассматриваемых в книге задач можно познакомиться в следующих руководствах:

Владимиров В. С. Уравнения математической физики. — М.: Наука, 1971.

Годунов С. К. Уравнения математической физики. — М.: Наука, 1971.

Соболев С. Л. Уравнения математической физики. — М.: Наука, 1966.

Тихонов А. Н., Самарский А. А. Уравнения математической физики. — М.: Наука, 1966.

2. Следующие работы посвящены в основном теоретическим аспектам метода конечных элементов:

Деклу Ж. Метод конечных элементов. Пер. с франц. — М.: Мир, 1976.

Корнеев В. Г. Схемы метода конечных элементов высоких порядков точности. — Л., — Издательство ЛГУ, 1977.

Оганесян Л. А., Ривкинд В. Я., Руховец Л. А. Вариационно-разностные методы решения эллиптических уравнений. В сб. «Дифференциальные

уравнения и их приложения». — Вильнюс: (часть 1 в вып. 5, 1973, часть 2 в вып. 8, 1974).

Сьярле Р. Метод конечных элементов для эллиптических задач. Пер. с франц. — М.: Мир, 1980.

3. Следующие работы ориентированы на конкретные приложения:

Квитка А. Л., Ворошко П. П., Бобрицкая С. Д. Напряженно-деформированное состояние тел вращения. — Киев: Наукова думка, 1977.

Коннор Дж., Бреббиа К. Метод конечных элементов в механике жидкости. Пер. с англ. — М.: Судостроение, 1979.

Методы расчета стержневых систем, пластин и оболочек с использованием ЭЦВМ, ч. 1, 2. Сб. под ред. А. Ф. Смирнова. — М.: Стройиздат, 1976.

Постнов В. А., Хархурим И. Я. Метод конечных элементов в расчетах судовых конструкций. — М.: Судостроение, 1974.

Постнов В. А. Численные расчеты судовых конструкций. — Л.: Судостроение, 1977.

Розин Л. А. Основы метода конечных элементов в теории упругости. — Л.: Издательство ЛПИ, 1972.

Розин Л. А. Метод конечных элементов в применении к упругим системам. — М.: Стройиздат, 1977.

Синицын А. Н. Метод конечных элементов в динамике сооружений. — М.: Стройиздат, 1978.

Слесарев И. С., Сиротин А. М. Вариационно-разностные схемы в теории переноса нейтронов. — М.: Атомиздат, 1978.

4. Более подробное изложение вариационных принципов механики и их связи с методом конечных элементов можно найти в упомянутой выше книге В. А. Постнова, а также в книге

Абовский Н. П., Андреев Н. Т., Деруга А. П. Вариационные принципы теории упругости и теории оболочек. — М.: Наука, 1978.

5. В книге почти не затронуты вопросы о численных методах решения получаемых линейных систем. Представление об этих методах можно получить по следующим работам:

Дафф И. С. Обзор исследований по разреженным матрицам. ТИИЭР, т. 65, № 4, 1977.

Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений. — М.: Наука, 1978.

Тьюарсон Р. Разреженные матрицы. Пер. с англ. — М.: Мир, 1977.

6. В третьей и шестой главах книги описывается метод переменных направлений Галеркина. Это отражает тот факт, что и в рамках метода конечных элементов стоит проблема создания экономичных схем. Большой материал по построению различных схем расщепления можно найти в следующих работах:

Марчук Г. И. Методы вычислительной математики. — М.: Наука, 1977.

Самарский А. А. Теория разностных схем. — М.: Наука, 1977.

Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики. — Новосибирск, 1967.

7. Следующие сборники статей дают некоторое представление о том, над чем работают советские специалисты, использующие метод конечных элементов:

Вариационно-разностные методы в математической физике. Сборник научных трудов. — Новосибирск: ВЦ СО АН СССР, 1974.

Вариационно-разностные методы в математической физике. Материалы Всесоюзной конференции. — Новосибирск: ВЦ СО АН СССР, 1978.

Вариационно-разностные методы решения задач математической физики. Труды 2-го Всесоюзного семинара. — Новосибирск: ВЦ СО АН СССР, 1976.

Метод конечных элементов в строительной механике. — Горький: Издательство ГГУ, 1975.

Разностные и вариационно-разностные методы. Труды семинара «Методы вычислительной и прикладной математики», вып. 2. — Новосибирск: ВЦ СО АН СССР, 1977.

8. Замечания об истории возникновения и развития метода конечных элементов можно найти в указанной выше книге А. Н. Синицына, а также в обзорной статье:

Зенкевич О. Метод конечных элементов: от интуиции к общности. Механика (сб. переводов), № 6, 1970.

ОГЛАВЛЕНИЕ

Э. МИТЧЕЛЛ, Р. УЭЙТ

МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Научный редактор А. А. Бряндинская
Мл. научный редактор И. К. Недобой
Художник В. В. Кашлин
Художественный редактор В. И. Шаповалов
Технический редактор Л. П. Чуркина
Корректор Н. В. Андреева

ИБ № 2219

Сдано в набор 01.04.80. Подписано к печати 23.12.80. Формат 60×90¹/₁₆ Бумага типографская № 2. Гарнитура латинская. Печать высокая. Объем 6,75 бум. л. Усл. печ. л. 13,50. Уч.-изд. л. 11,52. Изд № 1/0708. Тираж 12 000 экз. Зак. 672. Цена 85 коп.

ИЗДАТЕЛЬСТВО „МИР"
Москва, 1-й Рижский пер., 2

Ленинградская типография № 2 головное предприятие ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 198052, г. Ленинград, Л-52, Измайловский проспект, 29.

85 коп.